Appendix for DeltaGrad: Rapid retraining of machine learning models

Yinjun Wu, Edgar Dobriban, Susan B. Davidson

Contents

A Mathematical details

The main result for DeltaGrad with GD is Theorem 5, proved in Section A.2.8.

A.1 Additional notes on setup, preliminaries

A.1.1 Classical results on GD convergence, SGD convergence

Lemma 1 (GD convergence, folklore, e.g., Boyd and Vandenberghe (2004))**.** *Gradient descent over a strongly convex objective function with fixed step size $\eta_t = \eta \leq \frac{2}{L+\mu}$ has exponential convergence rate, i.e.:*

$$F(\boldsymbol{w}_t) - F(\boldsymbol{w}^*) \leq c^t \frac{L}{2} ||\boldsymbol{w}_0 - \boldsymbol{w}^*||^2 \tag{1}$$

where $c := (L-\mu)/(L+\mu) < 1$.

Recall also that the eigenvalues of the "contraction operator" $\mathbf{I} - \eta_t \mathbf{H}(\mathbf{w})$ are bounded as follows.

Lemma 2 (Classical bound on eigenvalues of the "contraction operator")**.** *Under the convergence conditions of gradient descent with fixed step size, i.e. $\eta_t = \eta \leq \frac{2}{\mu+L}$, the following inequality holds for any parameter $\boldsymbol{w}$:*

$$\|\boldsymbol{I} - \eta \boldsymbol{H}(\boldsymbol{w})\| \leq 1. \tag{2}$$

This lemma follows directly, because the eigenvalues of $I - \eta\mathbf{H}$ are bounded between $-1 \leq 1 - \eta L \leq 1 - \eta\mu \leq 1$.

Lemma 3 (SGD convergence, see e.g., Bottou et al. (2018))**.** *Suppose that the stochastic gradient estimates are correlated with the true gradient, and bounded in the following way. There exist two scalars $J_1 \geq J_2 > 0$ such that for arbitrary $\mathcal{B}_t$, the following two inequalities hold:*

$$\nabla F(\boldsymbol{w}_t)^T \mathbb{E} \frac{1}{B_t} \sum_{i \in \mathcal{B}_t} \nabla F_i(\boldsymbol{w}_t) \geq J_2 \|\nabla F(\boldsymbol{w}_t)\|^2, \tag{3}$$

$$\|\mathbb{E} \frac{1}{B_t} \sum_{i \in \mathcal{B}_t} \nabla F_i(\boldsymbol{w}_t)\| \leq J_1 \|\nabla F(\boldsymbol{w}_t)\|.$$

Also, assume that for two scalars $J_3, J_4 \geq 0$, we have:

$$Var\left(\frac{1}{B_t} \sum_{i \in \mathcal{B}_t} \nabla F_i(\boldsymbol{w}_t)\right) \leq J_3 + J_4 \|\nabla F(\boldsymbol{w}_t)\|^2. \tag{4}$$

By combining equations (3)-(4), the following inequality holds:

$$\mathbb{E}\|\frac{1}{B_t} \sum_{i \in \mathcal{B}_t} \nabla F_i(\boldsymbol{w}_t)\|^2 \leq J_3 + J_5 \|\nabla F(\boldsymbol{w}_t)\|^2$$

where $J_5 = J_4 + J_1^2 \geq J_2^2 \geq 0$.
Then stochastic gradient descent with fixed step size $\eta_t = \eta \leq \frac{J_2}{LJ_5}$ has the convergence rate:

$$\mathbb{E}[F(\boldsymbol{w}_t) - F(\boldsymbol{w}^*)] \leq \frac{\eta L J_3}{2\mu J_2} + (1 - \eta\mu J_2)^{t-1} \left(F(\boldsymbol{w}_1) - F(\boldsymbol{w}^*) - \frac{\eta L J_3}{2\mu J_2}\right) \to \frac{\eta L J_3}{2\mu J_2}.$$

If the gradient estimates are unbiased, then $\mathbb{E}\frac{1}{B_t}\sum_{i\in\mathcal{B}_t} \nabla F_i(\mathbf{w}_t) = \frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{w}_t) = \nabla F(\mathbf{w}_t)$ and thus $J_1 = J_2 = 1$. Moreover, $J_3 \sim 1/m$, where m is the minibatch size, because J_2 is the variance of the stochastic gradient.

So the convergence condition for fixed step size becomes $\eta_t = \eta \leq \frac{1}{LJ_5}$, in which $J_5 = J_4 + J_1^2 = J_4 + 1 \geq 1$. So $\eta_t = \eta \leq \frac{1}{LJ_5} \leq \frac{1}{L}$ suffices to ensure convergence.

A.1.2 Notations for DeltaGrad with SGD

The SGD parameters trained over the full dataset, explicitly trained over the remaining dataset and incrementally trained over the remaining dataset are denoted by $\mathbf{w}^S$, $\mathbf{w}^{U,S}$ and $\mathbf{w}^{I,S}$ respectively. Then given the mini-batch size B, mini-batch $\mathcal{B}_t$, the number of removed samples from each mini-batch ΔB_t and the set of removed samples R, the update rules for the three parameters are:

$$\mathbf{w}^S{}_{t+1} = \mathbf{w}^S{}_t - \eta \frac{1}{B} \sum_{i \in \mathcal{B}_t} \nabla F_i(\mathbf{w}^S{}_t) = \mathbf{w}^S{}_t - \eta G_{B,S}(\mathbf{w}^S{}_t), \tag{5}$$

$$\begin{aligned} \mathbf{w}^{U,S}{}_{t+1} &= \mathbf{w}^{U,S}{}_t - \eta \frac{1}{B - \Delta B_t} \sum_{i \in \mathcal{B}_t, i \notin R} \nabla F_i(\mathbf{w}^{U,S}{}_t) \\ &= \mathbf{w}^{U,S}{}_t - \eta G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_t), \end{aligned} \tag{6}$$

$$\mathbf{w}^{I,S}{}_{t+1} = \begin{cases} \mathbf{w}^{I,S}{}_t - \frac{\eta}{B-\Delta B_t} \sum_{i \in \mathcal{B}_t, i \notin R} \nabla F(\mathbf{w}^{I,S}{}_t) & (t - j_0) \mod T_0 = 0 \\ & \text{or } t \leq j_0 \\ \mathbf{w}^{I,S}{}_t - \frac{\eta}{B-\Delta B_t} \{B[\mathbf{B}_{j_m}(\mathbf{w}^{I,S}{}_t - \mathbf{w}^S{}_t) + \frac{1}{B} \sum_{i \in \mathcal{B}_t} \nabla F_i(\mathbf{w}^S{}_t)] - & otherwise \\ \sum_{i \in R, i \in \mathcal{B}_t} \nabla F(\mathbf{w}^{I,S}{}_t)\} & \end{cases} \tag{7}$$

in which $G_{B,S}(\mathbf{w}^S{}_t)$ and $G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_t)$ represent the average gradients over the minibatch $\mathcal{B}_t$ before and after removing samples.

We assume that the minibatch randomness of $\mathbf{w}^{U,S}$ and $\mathbf{w}^{I,S}$ is the same as $\mathbf{w}^S$. By following Lemma 3, we assume that the gradient estimates of SGD are unbiased, i.e. $\mathbb{E}\left(\frac{1}{B_t} \sum_{i \in \mathcal{B}_t} \nabla F_i(\mathbf{w})\right) = \frac{1}{n} \sum_{i=1}^n \nabla F_i(\mathbf{w}) = \nabla F(\mathbf{w})$ for any $\mathbf{w}$, which indicates that:

$$\begin{aligned} \mathbb{E}\left(\frac{1}{B} \sum_{i \in \mathcal{B}_t} \nabla F_i(\mathbf{w}^S{}_t)\right) &= \frac{1}{n} \sum_{i=1}^n \nabla F_i\left(\mathbf{w}^S{}_t\right) = \nabla F\left(\mathbf{w}^S{}_t\right), \\ \mathbb{E}\left(\frac{1}{B - \Delta B_t} \sum_{i \in \mathcal{B}_t, i \notin R} \nabla F_i(\mathbf{w}^{U,S}{}_t)\right) &= \frac{1}{n - \Delta n} \sum_{i \notin R} \nabla F_i\left(\mathbf{w}^{U,S}{}_t\right) = \nabla F^U\left(\mathbf{w}^{U,S}{}_t\right). \end{aligned}$$

A.1.3 Classical results for random variables

To analyze DeltaGrad with SGD, Bernstein's inequality (Oliveira, 2009; Tropp, 2012, 2016, e.g.,) is necessary. Both its scalar version and matrix version are stated below.

Lemma 4 (Bernstein's inequality for scalars)**.** *Consider a list of independent random variables, $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_k$ satisfying $\mathbb{E}(\boldsymbol{S}_i) = \boldsymbol{0}$ and $|\boldsymbol{S}_i| \leq J$, and their sum $\boldsymbol{Z} = \sum_{i=1}^k \boldsymbol{S}_i$. Then the following inequality holds:*

$$Pr(\|\boldsymbol{Z}\| \geq x) \leq \exp\left(\frac{-x^2}{\sum_{i=1}^k \mathbb{E}(\boldsymbol{S}_i^2) + \frac{Jx}{3}}\right), \forall x \geq 0.$$

Lemma 5 (Bernstein's inequality for matrices)**.** *Consider a list of independent $d_1 \times d_2$ random matrices, $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_k$ satisfying $E(\boldsymbol{S}_i) = \boldsymbol{0}$ and $\|\boldsymbol{S}_i\| \leq J$, and their sum $\boldsymbol{Z} = \sum_{i=1}^k \boldsymbol{S}_i$. Define the deterministic "varianc surrogate":*

$$V(\boldsymbol{Z}) = \max\left(\|\sum_{i=1}^k \mathbb{E}(\boldsymbol{S}_i \boldsymbol{S}_i^*)\|, \|\sum_{i=1}^k \mathbb{E}(\boldsymbol{S}_i^* \boldsymbol{S}_i)\|\right). \tag{8}$$

Then the following inequalities hold:

$$Pr(\|\boldsymbol{Z}\| \geq x) \leq (d_1 + d_2) \exp\left(\frac{-x^2}{V(\boldsymbol{Z}) + \frac{Jx}{3}}\right), \forall x \geq 0, \tag{9}$$

$$\mathbb{E}(\|\boldsymbol{Z}\|) \leq \sqrt{2V(\boldsymbol{Z}) \log(d_1 + d_2)} + \frac{1}{3} J \log(d_1 + d_2). \tag{10}$$

A.2 Results for deterministic gradient descent

The main result for DeltaGrad with GD is Theorem 5, proved in Section A.2.8.

A.2.1 Quasi-Newton

By following equations 1.2 and 1.3 in Byrd et al. (1994), the Quasi-Hessian update can be written as:

$$\mathbf{B}_{t+1} = \mathbf{B}_t - \frac{\mathbf{B}_t \Delta w_t \Delta w_t^T \mathbf{B}_t}{\Delta w_t^T \mathbf{B}_t \Delta w_t} + \frac{\Delta g_t \Delta g_t^T}{\Delta g_t^T \Delta w_t}. \tag{11}$$

We have used the indices k to index the Quasi-Hessians $\mathbf{B}_{j_k}$. This allows us to see that they correspond to the appropriate parameter gap Δw_{j_k} and gradient gap Δg_{j_k}. The indices j_k *depend on the iteration number* t in the main algorithm, and they are updated by removing the "oldest" entry, and adding T_0 at every period.

DeltaGrad uses equation (11) on the prior updates:

$$\mathbf{B}_{j_{k+1}} = \mathbf{B}_{j_k} - \frac{\mathbf{B}_{j_k} \Delta w_{j_k} \Delta w_{j_k}^T \mathbf{B}_{j_k}}{\Delta w_{j_k}^T \mathbf{B}_{j_k} \Delta w_{j_k}} + \frac{\Delta g_{j_k} \Delta g_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}, \tag{12}$$

where the initialized matrix $\mathbf{B}_{j_0}$ is $\mathbf{B}_{j_0} = \Delta g_{i_0}^T \Delta w_{j_0} \, / [\Delta w_{i_0}^T \Delta w_{j_0}] \mathbf{I}$.

We use formulas 3.5 and 2.25 from Byrd et al. (1994) for the Quasi-Newton method, with the caveat that they use slightly different notation.

For the update rule of $\mathbf{B}_{j_k}$, i.e.:

$$\mathbf{B}_{j_{k+1}} = \mathbf{B}_{j_k} - \frac{\mathbf{B}_{j_k} \Delta w_{j_k} \Delta w_{j_k}^T \mathbf{B}_{j_k}}{\Delta w_{j_k}^T \mathbf{B}_{j_k} \Delta w_{j_k}} + \frac{\Delta g_{j_k} \Delta g_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}. \tag{13}$$

There is an equivalent expression for the inverse of $\mathbf{B}_{j_k}$ as below:

$$\mathbf{B}_{j_{k+1}}^{-1} = \left(\mathbf{I} - \frac{\Delta w_{j_k} \Delta g_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\right) \mathbf{B}_{j_k}^{-1} \left(\mathbf{I} - \frac{\Delta g_{j_k} \Delta w_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\right) + \frac{\Delta w_{j_k} \Delta w_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}. \tag{14}$$

See Algorithm 1 for an overview of the L-BFGS algorithm.

Algorithm 1: Overview of L-BFGS algorithm

Input : The sequence of the model parameter differences $\Delta W = \{\Delta w_0, \Delta w_1, \ldots, \Delta w_{m-1}\}$, the sequence of the gradient differences $\Delta G = \{\Delta g_0, \Delta g_1, \ldots, \Delta g_{m-1}\}$, a vector $\mathbf{v}$, history size m

Output: Approximate results of $\mathbf{H}(w_m)\mathbf{v}$ at point w_m, and for some $\mathbf{v}$, such that $\Delta w_i \approx w_i - w_{i-1}$ for all i

Compute $\Delta W^T \Delta W$

Compute $\Delta W^T \Delta G$, get its diagonal matrix D and its lower triangular submatrix L

Compute $\sigma = \Delta g_{m-1}^T \Delta w_{m-1} / \left(\Delta w_{m-1}^T \Delta w_{m-1}\right)$

Compute the Cholesky factorization for $\sigma \Delta W^T \Delta W + L D L^T$ to get $J J^T$

Compute $p = \begin{bmatrix} -D^{\frac{1}{2}} & D^{-\frac{1}{2}} L^T \\ \mathbf{0} & J^T \end{bmatrix}^{-1} \begin{bmatrix} D^{\frac{1}{2}} & \mathbf{0} \\ D^{-\frac{1}{2}} L^T & J^T \end{bmatrix}^{-1} \begin{bmatrix} \Delta G^T \mathbf{v} \\ \sigma \Delta W^T \mathbf{v} \end{bmatrix}$

return $\sigma \mathbf{v} - \begin{bmatrix} \Delta G & \sigma \Delta W \end{bmatrix} p$

A.2.2 Proof that Quasi-Hessians are well-conditioned

We show that the Quasi-Hessian matrices computed by L-BFGS are well-conditioned.

Lemma 6 (Bounds on Quasi-Hessians)**.** *The Quasi-Hessian matrices $\boldsymbol{B}_{j_k}$ are well-conditioned. There exist two positive constants K_1 and K_2 (depending on the problem parameters μ, L, etc) such that for any t, any vector $\boldsymbol{z}$, and all $k \in \{0, 1, \ldots, m\}$, the following inequality holds:*

$$K_1\|\boldsymbol{z}\|^2 \leq \boldsymbol{z}^T \boldsymbol{B}_{j_k} \boldsymbol{z} \leq K_2\|\boldsymbol{z}\|^2.$$

Proof. We start with the lower bound. Based on equation (14), $\|\mathbf{B}_{j_k}^{-1}\|$ can be bounded by:

$$\|\mathbf{B}_{j_{k+1}}^{-1}\| \leq \|\mathbf{I} - \frac{\Delta w_{j_k} \Delta g_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\| \cdot \|\mathbf{B}_{j_k}^{-1}\| \cdot \|\mathbf{I} - \frac{\Delta g_{j_k} \Delta w_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\| + \|\frac{\Delta w_{j_k} \Delta w_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\|. \tag{15}$$

in which by using the mean value theorem, $\|\mathbf{I} - \frac{\Delta w_{j_k} \Delta g_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\|$ can be bounded as:

$$\begin{aligned} &\|\mathbf{I} - \frac{\Delta w_{j_k} \Delta g_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\| \leq 1 + \frac{\|\Delta w_{j_k} \Delta g_{j_k}^T\|}{\Delta g_{j_k}^T \Delta w_{j_k}} \\ &= 1 + \frac{\|\Delta w_{j_k} (\mathbf{H}_{j_k} \Delta w_{j_k})^T\|}{\Delta w_{j_k}^T \mathbf{H}_{j_k} \Delta w_{j_k}} \leq 1 + \frac{\|\Delta w_{j_q}\| \|\mathbf{H}_{j_k}\| \|\Delta w_{j_q}\|}{\mu \|\Delta w_{j_q}\|^2} \leq 1 + \frac{L}{\mu}. \end{aligned} \tag{16}$$

In addition, $\|\frac{\Delta w_{j_k} \Delta w_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\|$ can be bounded as:

$$\|\frac{\Delta w_{j_k} \Delta w_{j_k}^T}{\Delta g_{j_k}^T \Delta w_{j_k}}\| = \|\frac{\Delta w_{j_k} \Delta w_{j_k}^T}{\Delta w_{j_k}^T \mathbf{H}_{j_k} \Delta w_{j_k}}\| \leq \|\frac{\Delta w_{j_k}^T \Delta w_{j_k}}{\mu \Delta w_{j_k}^T \Delta w_{j_k}}\| = \frac{1}{\mu}. \tag{17}$$

So by combining Equation (16) and Equation (17), Equation (15) can be bounded by:

$$\begin{aligned} \|\mathbf{B}_{j_{k+1}}^{-1}\| &\leq (1 + \frac{L}{\mu})^2 \|\mathbf{B}_{j_k}^{-1}\| + \frac{1}{\mu} \leq (1 + \frac{L}{\mu})^{2k} \|\mathbf{B}_{j_0}^{-1}\| + \frac{1 - (1 + \frac{L}{\mu})^{2k}}{1 - (1 + \frac{L}{\mu})^2} \frac{1}{\mu} \\ &= (1 + \frac{L}{\mu})^{2k} \frac{L}{\mu} + \frac{1 - (1 + \frac{L}{\mu})^{2k}}{1 - (1 + \frac{L}{\mu})^2} \frac{1}{\mu}. \end{aligned}$$

which thus implies that $\|\mathbf{B}_{j_k}\| \geq K_1 := \frac{1}{(1+\frac{L}{\mu})^{2k} \frac{L}{\mu} + \frac{1-(1+\frac{L}{\mu})^{2k}}{1-(1+\frac{L}{\mu})^2} \frac{1}{\mu}}$ where $0 \leq k \leq m$. Recall that m is small, (set as $m = 2$ in the experiments). So the lower bound will not approach zero.

Then based on Equation (11), we derive an upper bound for $\|\mathbf{B}_{j_k}\|$ as follows:

$$\begin{aligned} \mathbf{z}^T \mathbf{B}_{j_{k+1}} \mathbf{z} &= \mathbf{z}^T \mathbf{B}_{j_k} \mathbf{z} - \frac{\mathbf{z}^T \mathbf{B}_{j_k} \Delta w_{j_k} \Delta w_{j_k}^T \mathbf{B}_{j_k} \mathbf{z}}{\Delta w_{j_k}^T \mathbf{B}_{j_k} \Delta w_{j_k}} + \frac{\mathbf{z}^T \Delta g_{j_k} \Delta g_{j_k}^T \mathbf{z}}{\Delta g_{j_k}^T \Delta w_{j_k}} \\ &\leq \mathbf{z}^T \mathbf{B}_{j_k} \mathbf{z} + \frac{\mathbf{z}^T \Delta g_{j_k} \Delta g_{j_k}^T \mathbf{z}}{\Delta g_{j_k}^T \Delta w_{j_k}} = \mathbf{z}^T \mathbf{B}_{j_k} \mathbf{z} + \frac{\mathbf{z}^T \mathbf{H}_{j_k} \Delta w_{j_k} \Delta w_{j_k}^T \mathbf{H}_{j_k} \mathbf{z}}{\Delta w_{j_k}^T \mathbf{H}_{j_k} \Delta w_{j_k}} \\ &\leq \mathbf{z}^T \mathbf{B}_{j_k} \mathbf{z} + \frac{\mathbf{z}^T \mathbf{H}_{j_k} \mathbf{z} \Delta w_{j_k}^T \mathbf{H}_{j_k} \Delta w_{j_k}}{\Delta w_{j_k}^T \mathbf{H}_{j_k} \Delta w_{j_k}} = \mathbf{z}^T \mathbf{B}_{j_k} \mathbf{z} + \mathbf{z}^T \mathbf{H}_{j_k} \mathbf{z} \\ &\leq \mathbf{z}^T \mathbf{B}_{j_k} \mathbf{z} + L\|\mathbf{z}\|^2. \end{aligned}$$

The first inequality uses the fact that $\mathbf{z}^T \mathbf{B}_{j_k} \Delta w_{j_k} \Delta w_{j_k}^T \mathbf{B}_{j_k} \mathbf{z} = \left(\mathbf{z}^T \mathbf{B}_{j_k} \Delta w_{j_k}\right)^2 \geq 0$ and $\Delta w_{j_k}^T \mathbf{B}_{j_k} \Delta w_{j_k} \geq 0$, due to the positive definiteness of $\mathbf{B}_{j_k}$. The second inequality uses the Cauchy-Schwarz inequality for the Quasi-Hessian, i.e.:

$$\left(\mathbf{a}^T \mathbf{H}_{j_k} \mathbf{b}\right)^2 \leq \left(\mathbf{a}^T \mathbf{H}_{j_k} \mathbf{a}\right) \left(\mathbf{b}^T \mathbf{H}_{j_k} \mathbf{b}\right).$$

By applying the formula above recursively, we get $\mathbf{z}^T \mathbf{B}_{j_{k+1}} \mathbf{z} \leq (k+1)L\|\mathbf{z}\|^2$ where $0 \leq k \leq m$. Again, as m is bounded, so we have $(k+1)L \leq K_2 := (m+1)L$. This finishes the proof. □

A.2.3 Proof preliminaries

First of all, we provide the bound on δ_t, which is defined as:

Lemma 7 (Upper bound on δ_t)**.** *By defining*

$$\delta_t = -\frac{\eta}{n-r}\left(\frac{r}{n}\sum_{i=1}^{n}\nabla F_i\left(\boldsymbol{w}^U{}_t\right) - \sum_{i\in R}\nabla F_i\left(\boldsymbol{w}^U{}_t\right)\right),$$

we then have $\|\delta_t\| \leq 2c_2\frac{r\eta}{n}$.

Proof. Based on the definition of δ_t, we can rearrange it a little bit as:

$$\begin{aligned}
\|\delta_t\| &= \|-\frac{\eta r}{n\,(n-r)}\sum_{i=1}^{n}\nabla F_i\left(\mathbf{w}^U{}_t\right) + \frac{\eta}{n-r}\sum_{i\in R}\nabla F_i\left(\mathbf{w}^U{}_t\right)\| \\
&= \|-\frac{\eta r}{n\,(n-r)}[\sum_{i=1}^{n}\nabla F_i\left(\mathbf{w}^U{}_t\right) - \sum_{i\in R}\nabla F_i\left(\mathbf{w}^U{}_t\right)] + (\frac{\eta}{n-r} - \frac{\eta r}{n(n-r)})\sum_{i\in R}\nabla F_i\left(\mathbf{w}^U{}_t\right)\| \\
&= \|-\frac{\eta r}{n\,(n-r)}\sum_{i\notin R}\nabla F_i\left(\mathbf{w}^U{}_t\right) + \frac{\eta}{n}\sum_{i\in R}\nabla F_i\left(\mathbf{w}^U{}_t\right)\|.
\end{aligned}$$

Then by using the triangle inequality and Assumption 3 (bounded gradients), the formula above can be bounded as:

$$\leq \frac{\eta r}{n\,(n-r)}\sum_{i\notin R}\|\nabla F_i\left(\mathbf{w}^U{}_t\right)\| + \frac{\eta}{n}\sum_{i\in R}\|\nabla F_i\left(\mathbf{w}^U{}_t\right)\| \leq \frac{\eta r}{n}c_2 + \frac{\eta r}{n}c_2 = \frac{2\eta r}{n}c_2$$

□

Notice that Algorithm 1 requires $2m$ vectors as the input, i.e. $[\Delta w_{j_0}, \Delta w_{j_1}, \ldots, \Delta w_{j_{m-1}}]$ and $[\Delta g_{j_0}, \Delta g_{j_1}, \ldots, \Delta g_{j_{m-1}}]$ to approximate the product of the Hessian matrix $\mathbf{H}(w_t)$ and the input vector Δw_t at the t_{th} iteration where $j_{m-1} \leq t \leq j_{m-1} + T_0$.

Note that by multiplying Δw_{j_k} on both sides of the Quasi-Hessian update Equation (12), we have the classical *secant equation* that characterizes Quasi-Newton methods as below:

$$\mathbf{B}_{j_{k+1}}\Delta w_{j_k} = \Delta g_{j_k}. \tag{18}$$

Then we give a bound on the quantity $\|\Delta g_{j_k} - \mathbf{B}_{j_q}\Delta w_{j_k}\|$ where the intermediate index q is in between the "correct" index $k+1$ and the final index m, so $m \geq q \geq k+1$. This characterizes the error by using a different Quasi-Hessian at some iteration. Its proof borrows ideas from Conn et al. (1991). Unlike Conn et al. (1991), our proof relies on a preliminary estimate on the bound on $\|\mathbf{w}_t - \mathbf{w}^I{}_t\|$, which is at the level of $O(\frac{r}{n})$. The proof of the bound will be presented later.

Theorem 1. *Suppose that the preliminary estimate is:* $\|\boldsymbol{w}_{j_k} - \boldsymbol{w}^I{}_{j_k}\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}$*, where* $k = 1, 2, \ldots, m$ *and* $M_1 = \frac{2c_2}{\mu}$*. Let* $e = \frac{L(L+1)+K_2L}{\mu K_1}$*, for the upper and lower bounds* K_1, K_2 *on the eigenvalues of the quasi-Hessian from Lemma 6, for the upper bounds* c_2 *on the gradient from Assumption 3 and for the Lipshitz constant* c_0 *of the Hessian. For* $1 \leq k+1 \leq q \leq m$*, we have:*

$$\|\boldsymbol{H}_{j_k} - \boldsymbol{H}_{j_q}\| \leq c_0 d_{j_k,j_q} + c_0\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}$$

and

$$\|\Delta g_{j_k} - \boldsymbol{B}_{j_q}\Delta w_{j_k}\| \leq \left[(1+e)^{q-k-1} - 1\right]\cdot c_0(d_{j_k,j_q} + \frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n})\cdot s_{j_1,j_m},$$

where $s_{j_1,j_m} = \max\left(\|\Delta w_a\|\right)_{a=j_1,j_2,\ldots,j_m}$ *and* d *is defined as the maximum gap between the steps of the algorithm over the iterations from* j_k *to* j_q*:*

$$d_{j_k,j_q} = \max\left(\|\boldsymbol{w}_a - \boldsymbol{w}_b\|\right)_{j_k\leq a\leq b\leq j_q}. \tag{19}$$

Proof. Let $v_q = \Delta g_{j_k} - \mathbf{B}_{j_{q+1}} \Delta w_{j_k}$, $b_q = \|v_q\|$ and $f = c_0(d_{j_1, j_m+T_0-1} + \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}) s_{j_1,j_m}$.

Let us bound the difference between the averaged Hessians $\|\mathbf{H}_{j_k} - \mathbf{H}_{j_q}\|$, where $1 \le k < q \le m$, using their definition, as well as using Assumption 4 on the Lipshitzness of the Hessian:

$$\begin{aligned}
&\|\mathbf{H}_{j_k} - \mathbf{H}_{j_q}\| \\
&= \|\int_0^1 [\mathbf{H}(\mathbf{w}_{j_k} + x(\mathbf{w}^I{}_{j_k} - \mathbf{w}_{j_k}))]dx - \int_0^1 [\mathbf{H}(\mathbf{w}_{j_q} + x(\mathbf{w}^I{}_{j_q} - \mathbf{w}_{j_q}))]dx\| \\
&= \|\int_0^1 [\mathbf{H}(\mathbf{w}_{j_k} + x(\mathbf{w}^I{}_{j_k} - \mathbf{w}_{j_k})) - \mathbf{H}(\mathbf{w}_{j_q} + x(\mathbf{w}^I{}_{j_q} - \mathbf{w}_{j_q}))]dx\| \\
&\le c_0 \int_0^1 \|\mathbf{w}_{j_k} + x(\mathbf{w}^I{}_{j_k} - \mathbf{w}_{j_k}) - [\mathbf{w}_{j_q} + x(\mathbf{w}^I{}_{j_q} - \mathbf{w}_{j_q})]\| dx \\
&\le c_0 \|\mathbf{w}_{j_k} - \mathbf{w}_{j_q}\| + \frac{c_0}{2} \|\mathbf{w}^I{}_{j_k} - \mathbf{w}_{j_k} - (\mathbf{w}^I{}_{j_q} - \mathbf{w}_{j_q})\| \\
&\le c_0 \|\mathbf{w}_{j_k} - \mathbf{w}_{j_q}\| + \frac{c_0}{2} \|\mathbf{w}_{j_q} - \mathbf{w}^I{}_{j_q}\| + \frac{c_0}{2} \|\mathbf{w}^I{}_{j_k} - \mathbf{w}_{j_k}\| \\
&\le c_0 d_{j_k,j_q} + \frac{c_0}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n} \le c_0 d_{j_1,j_m+T_0-1} + \frac{c_0}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}.
\end{aligned} \tag{20}$$

On the last line, we used the definition of d_{j_k,j_q}, and the assumption on the boundedness of $\|\mathbf{w}^I{}_{j_k} - \mathbf{w}_{j_k}\|$.

Then, when $q = k$, according to Equation (18), the secant equation $\Delta g_{j_k} = \mathbf{B}_{j_{k+1}} \Delta w_{j_k}$ holds. So $\|\Delta g_{j_k} - \mathbf{B}_{j_{k+1}} \Delta w_{j_k}\| = 0$, which proves the claim when $q = k$. So $v_q = b_q = 0$.

Next, let $u_q = \Delta g_{j_q} - \mathbf{B}_{j_q} \Delta w_{j_q}$. This quantity is closely related to $v_{q-1} = \Delta g_{j_k} - \mathbf{B}_{j_q} \Delta w_{j_k}$, and the difference is that in u_q, the $\Delta g, \Delta w$ terms are defined at q, as opposed to the base one at k. Then $|u_q^T \Delta w_{j_k}|$, where $q > k$, can be bounded as:

$$\begin{aligned}
&|u_q^T \Delta w_{j_k}| \\
&= |\Delta g_{j_q}^T \Delta w_{j_k} - \Delta g_{j_k}^T \Delta w_{j_q} + \Delta g_{j_k}^T \Delta w_{j_q} - \Delta w_{j_q}^T \mathbf{B}_{j_q} \Delta w_{j_k}| \\
&\le |\Delta g_{j_q}^T \Delta w_{j_k} - \Delta g_{j_k}^T \Delta w_{j_q}| + |\Delta w_{j_q}^T v_{q-1}| \\
&\le |\Delta g_{j_q}^T \Delta w_{j_k} - \Delta g_{j_k}^T \Delta w_{j_q}| + \|\Delta w_{j_q}\| \cdot b_{q-1} \\
&= |\Delta w_{j_q}^T \mathbf{H}_{j_q} \Delta w_{j_k} - \Delta w_{j_k}^T \mathbf{H}_{j_k} \Delta w_{j_q}| + \|\Delta w_{j_q}\| \cdot b_{q-1} \\
&= |\Delta w_{j_q}^T \left(\mathbf{H}_{j_q} - \mathbf{H}_{j_k}\right) \Delta w_{j_k}| + \|\Delta w_{j_q}\| \cdot b_{q-1} \\
&\le \|\Delta w_{j_q}\| \cdot \|\mathbf{H}_{j_q} - \mathbf{H}_{j_k}\| \cdot \|\Delta w_{j_k}\| + \|\Delta w_{j_q}\| \cdot b_{q-1} \\
&\le (f + b_{q-1}) \|\Delta w_{j_q}\|,
\end{aligned} \tag{21}$$

in which the first inequality uses the triangle inequality, the second inequality uses the Cauchy-Schwarz inequality, and the subsequent equality uses the Cauchy mean value theorem. Finally, the third inequality uses Assumption 4 and equation (20). We also use the following bounds, which hold by definition (notice that $k, q \le m$):

$$\|w_{j_k} - w_{j_q}\| \le d_{j_k,j_q} \qquad\qquad \|\Delta w_{j_q}\| \le s_{j_1,j_m}.$$

The argument on the upper bound of b_q will proceed by induction. The claim is true for the base case $q = k$. Assuming that the claim is true for $q-1$, we want to prove it for q, which is bounded as below:

$$b_q = \|\Delta g_{j_k} - \left(\mathbf{B}_{j_q} - \frac{\mathbf{B}_{j_q} \Delta w_{j_q} \Delta w_{j_q}^T \mathbf{B}_{j_q}}{\Delta w_{j_q}^T \mathbf{B}_{j_q} \Delta w_{j_q}} + \frac{\Delta g_{j_q} \Delta g_{j_q}^T}{\Delta g_{j_q}^T \Delta w_{j_q}}\right) \Delta w_{j_k}\|. \tag{22}$$

By using the triangle inequality, we obtain the following upper bound:

$$\le b_{q-1} + \|\left(\frac{\Delta g_{j_q} \Delta g_{j_q}^T}{\Delta g_{j_q}^T \Delta w_{j_q}} - \frac{\mathbf{B}_{j_q} \Delta w_{j_q} \Delta w_{j_q}^T \mathbf{B}_{j_q}}{\Delta w_{j_q}^T \mathbf{B}_{j_q} \Delta w_{j_q}}\right) \Delta w_{j_k}\|.$$

Now we come to a key and nontrivial step of the argument. By bringing fractions to the common denominator in the second term, adding and subtracting $\Delta g_{j_q}\Delta g_{j_q}^T\Delta w_{j_q}^T\Delta g_{j_q}$ and $\Delta g_{j_q}(\mathbf{B}_{j_q}\Delta w_{j_q})^T\Delta w_{j_q}^T\Delta g_{j_q}$, and rearranging to factor out the term $-u_q$ in the numerator of each summand, the formula above can be rewritten as:

$$= b_{q-1} + \frac{\|[-\Delta g_{j_q}\Delta g_{j_q}^T\Delta w_{j_q}^T u_q + \Delta g_{j_q}u_q^T\Delta w_{j_q}^T\Delta g_{j_q} + u_q\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_q}^T\Delta g_{j_q}]\Delta w_{j_k}\|}{\Delta g_{j_q}^T\Delta w_{j_q}\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_q}}.$$

Next, using the Cauchy mean value theorem, and the fact that the smallest eigenvalues of $\mathbf{H}_{j_q}, \mathbf{B}_{j_q}$ are lower bounded by μ, K_1 respectively, the formula above is bounded as:

$$\begin{aligned}
&\leq b_{q-1} + \frac{\|[-\Delta g_{j_q}\Delta g_{j_q}^T\Delta w_{j_q}^T u_q + \Delta g_{j_q}u_q^T\Delta w_{j_q}^T\Delta g_{j_q} + u_q\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_q}^T\Delta g_{j_q}]\Delta w_{j_k}\|}{\mu K_1\|\Delta w_{j_q}\|^4} \\
&\leq b_{q-1} + (\|\Delta g_{j_q}\|^2\cdot\|\Delta w_{j_q}^T u_q\Delta w_{j_k}\| + \|\Delta g_{j_q}\|\cdot\|u_q^T\Delta w_{j_q}^T\Delta g_{j_q}\Delta w_{j_k}\| \\
&+ \|u_q\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_k}\Delta w_{j_q}^T\Delta g_{j_q}\|)/\mu K_1\|\Delta w_{j_q}\|^4.
\end{aligned}$$

Now we want to bound the last three terms one by one. First of all, $\|\Delta g_{j_q}\|^2\|\Delta w_{j_q}^T u_q\Delta w_{j_k}\|$ can be bounded as:

$$\begin{aligned}
&\|\Delta g_{j_q}\|^2\cdot\|\Delta w_{j_q}^T u_q\Delta w_{j_k}\| = \|\mathbf{H}_{j_q}\Delta w_{j_q}\|^2\cdot|\Delta w_{j_q}^T u_q|\cdot\|\Delta w_{j_q}\| \\
&\leq L\|\Delta w_{j_q}\|^3\cdot|\Delta w_{j_q}^T u_q| \leq L\left(f + b_{q-1}\right)\|\Delta w_{j_q}\|^4,
\end{aligned}$$

in which the first equality uses the Cauchy mean value theorem, the subsequent inequality uses Assumption 3 and the last inequality uses equation (21), the upper bound on $|\Delta w_{j_q}^T u_q|$.

Then for $\|\Delta g_{j_q}\|\cdot\|u_q^T\Delta w_{j_q}^T\Delta g_{j_q}\Delta w_{j_k}\|$, we have a very similar argument. The only difference is that we factor out the scalar $\Delta w_{j_q}^T\Delta g_{j_q}$, and bound it by $L\|\Delta w_{j_q}\|^2$, i.e.:

$$\begin{aligned}
&\|\Delta g_{j_q}\|\cdot\|u_q^T\Delta w_{j_q}^T\Delta g_{j_q}\Delta w_{j_k}\| \\
&= \|\mathbf{H}_{j_q}\Delta w_{j_q}\|\cdot|\Delta w_{j_q}^T\Delta g_{j_q}|\cdot|u_q^T\Delta w_{j_k}| \\
&\leq L^2\left(f + b_{q-1}\right)\|\Delta w_{j_q}\|^4,
\end{aligned}$$

in which the first equality uses Cacuhy mean value theorem and the fact that $\Delta w_{j_q}^T\Delta g_{j_q}$ is a scalar and the last inequality uses Assumption 3 and Equation (21).

In terms of the bound on $\|u_q\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_k}\Delta w_{j_q}^T\Delta g_{j_q}\|$, it is derived as:

$$\begin{aligned}
&\|u_q\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_k}\Delta w_{j_q}^T\Delta g_{j_q}\| \\
&= \|u_q\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_k}\Delta w_{j_q}^T\Delta g_{j_q}\| \\
&\leq \|u_q\Delta w_{j_q}^T\|\cdot|\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_k}|\cdot\|\Delta g_{j_q}\| \\
&\leq \left(f + b_{q-1}\right)\|\Delta w_{j_q}\|\cdot|\Delta w_{j_q}^T\mathbf{B}_{j_q}\Delta w_{j_k}|\cdot\|\mathbf{H}_{j_q}\Delta w_{j_q}\| \\
&\leq \left(f + b_{q-1}\right)\|\Delta w_{j_q}\|\cdot K_2\|\Delta w_{j_q}\|^2\cdot L\|\Delta w_{j_q}\| \\
&= K_2 L\left(f + b_{q-1}\right)\|\Delta w_{j_q}\|^4,
\end{aligned}$$

in which the first inequality uses the Cauchy Schwarz inequality, the second inequality uses equation (21) and the third inequality uses Assumption 6.

In summary, for all $j \geq t+1$, Equation (22) is bounded by:

$$\begin{aligned}
b_q &\leq b_{q-1} + \frac{L(L+1) + K_2 L}{\mu K_1\|\Delta w_{j_q}\|^4}\left(f + b_{q-1}\right)\|\Delta w_{j_q}\|^4 \\
&= (1+e)b_{q-1} + ef.
\end{aligned}$$

By recursion and using the fact that $b_k = 0$, this can be bounded as:

$$\begin{aligned} &\leq (1+e)^{q-k} b_{k+1} + \sum_{i=0}^{q-k-1} (1+e)^i e \cdot f \\ &= \frac{(1+e)^{q-k}-1}{e} \cdot ef = [(1+e)^{q-k}-1]f. \end{aligned} \tag{23}$$

This proves the required claim $b_q \leq [(1+e)^{q-k}-1]f$ and finishes the proof. □

Corollary 1 (Approximation accuracy of quasi-Hessian to mean Hessian)**.** *Suppose that* $\|\boldsymbol{w}_{j_s} - \boldsymbol{w}^I{}_{j_s}\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$ *and* $\|\boldsymbol{w}_t - \boldsymbol{w}^I{}_t\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$ *where* $s = 1, 2, \ldots, m$*. Then for* $j_m \leq t \leq j_m + T_0 - 1$*,*

$$\|\boldsymbol{H}_t - \boldsymbol{B}_{j_m}\| \leq \xi_{j_1,j_m} := A d_{j_1,j_m+T_0-1} + A \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}, \tag{24}$$

where recall again that c_0 *is the Lipschitz constant of the Hessian,* d_{j_1,j_m+T_0-1} *is the maximal gap between the iterates of the GD algorithm on the full data from* j_1 *to* $j_m + T_0 - 1$ *(see equation* (19)*), which goes to zero as* $t \to \infty$*) and* $A = \frac{c_0\sqrt{m}[(1+e)^m-1]}{c_1} + c_0$ *in which* e *is a problem dependent constant defined in Theorem 1,* c_1 *is the "strong independence" constant from* (5)*.*

Proof. Based on Theorem 1, $b_{q-1} = \|\mathbf{H}_{j_q}\Delta w_{j_k} - \mathbf{B}_{j_q}\Delta w_{j_k}\| \leq \left[(1+e)^{q-k-1}-1\right] f$.

Then based on the "strong linear independence" in Assumption 5, the matrix $\Delta W_{j_1,j_2,\ldots,j_m} = [\frac{\Delta w_{j_1}}{s_{j_1,j_m}}, \frac{\Delta w_{j_2}}{s_{j_1,j_m}}, \ldots, \frac{\Delta w_{j_m}}{s_{j_1,j_m}}]$ has its smallest singular value lower bounded by $c_1 > 0$. Then $\|\mathbf{H}_{j_m} - \mathbf{B}_{j_m}\|$ can be bounded as below:

$$\begin{aligned} \|\mathbf{H}_{j_m} - \mathbf{B}_{j_m}\| &\leq \frac{1}{c_1} \|(\mathbf{H}_{j_m} - \mathbf{B}_{j_m})\, \Delta W_{j_1,j_2,\ldots,j_m}\| \\ &\leq \sqrt{m}[(1+e)^m - 1] \frac{c_0}{c_1} \left(d_{j_1,j_m+T_0-1} + \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n} \right) \end{aligned} \tag{25}$$

The second inequality uses the bound $\|M\| \leq \sqrt{m} \max_i \|m_i\|$, where M is a matrix with the m columns m_i.

So by combining the results from equation (25), we can upper bound $\|\mathbf{H}_t - \mathbf{B}_{j_m}\|$ where $j_m \leq t \leq j_m + T_0 - 1$, i.e.:

$$\begin{aligned} &\|\mathbf{H}_t - \mathbf{B}_{j_m}\| = \|\mathbf{H}_t - \mathbf{H}_{j_m} + \mathbf{H}_{j_m} + \mathbf{B}_{j_m}\| \\ &\leq \|\mathbf{H}_t - \mathbf{H}_{j_m}\| + \|\mathbf{H}_{j_m} - \mathbf{B}_{j_m}\| \\ &\leq c_0(d_{j_m,t} + M_1 \frac{r}{n}) + \sqrt{m}[(1+e)^m - 1] \frac{c_0}{c_1} \left(d_{j_1,j_m+T_0-1} + \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n} \right) \\ &\leq A d_{j_1,j_m+T_0-1} + A \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n} \end{aligned} \tag{26}$$

This finishes the proof. □

Note that in the upper bound on $\|\mathbf{H}_t - \mathbf{B}_{j_m}\|$, there is one term d_{j_1,j_m+T_0-1}. So we need to do some analysis of this term:

Lemma 8 (Contraction of the GD iterates)**.** *Recall the definition of* d_{j_k,j_q} *from Theorem 1:*

$$d_{j_k,j_q} = \max\left(\|\boldsymbol{w}_a - \boldsymbol{w}_b\|\right)_{j_k \leq a \leq b \leq j_q}.$$

Then $d_{j_k,j_q} \leq d_{j_k-z,j_q-z}$ *for any positive integers* z *and* $d_{j_k,j_q} \leq (1-\mu\eta)^{j_k} d_{0,j_q-j_k}$ *for any* $0 \leq j_k \leq j_q$*.*

Proof. To prove the two inequalities, we should look at d_{j_k,j_q} and d_{j_k-z,j_q-z} where z is a positive integer. For any given $j_k \leq a \leq b \leq j_q$, the upper bound on $\|w_a - w_b\|$ can be derived as below:

$$\begin{aligned}
&\|\mathbf{w}_a - \mathbf{w}_b\| = \|\mathbf{w}_{a-1} - \eta\nabla F(\mathbf{w}_{a-1}) - (\mathbf{w}_{b-1} - \eta\nabla F(\mathbf{w}_{b-1})\| \\
&= \|\mathbf{w}_{a-1} - \mathbf{w}_{b-1} - \eta(\nabla F(\mathbf{w}_{a-1}) - \nabla F(\mathbf{w}_{b-1}))\| \\
&= \|\mathbf{w}_{a-1} - \mathbf{w}_{b-1} - \\
&\eta\frac{1}{n}\left(\int_0^1 \sum_{i=1}^{n} \mathbf{H}_i(\mathbf{w}_{a-1} + x(\mathbf{w}_{b-1} - \mathbf{w}_{a-1}))dx\right)(\mathbf{w}_{a-1} - \mathbf{w}_{b-1})\| \\
&= \|\left(\mathbf{I} - \frac{\eta}{n}\left(\int_0^1 \sum_{i=1}^{n} \mathbf{H}_i(\mathbf{w}_{a-1} + x(\mathbf{w}_{b-1} - \mathbf{w}_{a-1}))dx\right)\right)(\mathbf{w}_{a-1} - \mathbf{w}_{b-1})\|.
\end{aligned}$$

The derivation above uses the update rule of gradient descent and Cauchy mean-value theorem. Then according to Cauchy Schwarz inequality and strong convexity, it can be further bounded as $\|\mathbf{w}_a - \mathbf{w}_b\| \leq (1 - \eta\mu)\|\mathbf{w}_{a-1} - \mathbf{w}_{b-1}\|$.

This can be used iteratively, which ends up with the following inequality:

$$\|\mathbf{w}_a - \mathbf{w}_b\| \leq (1 - \eta\mu)^z \|\mathbf{w}_{a-z} - \mathbf{w}_{b-z}\| \tag{27}$$

which indicates that $d_{j_k,j_q} \leq (1-\eta\mu)^z d_{j_k-z,j_q-z}$ and thus $d_{j_k,j_q} \leq d_{j_k-z,j_q-z}$. So by replacing z with j_k, we will have: $d_{j_k,j_q} \leq (1-\mu\eta)^{j_k} d_{0,j_q-j_k}$.

□

A.2.4 Main recursions

We bound the difference between $\mathbf{w}^I{}_t$ and $\mathbf{w}^U{}_t$. The proofs of the theorems stated below are in the following sections.

Our proof starts out with the usual approach of trying to show a contraction for the gradient updates, see e.g., Bottou et al. (2018). First we bound $\|\mathbf{w}_t - \mathbf{w}^U{}_t\|$, i.e.:

Theorem 2 (Bound between iterates on full and the leave-r-out dataset)**.** $\|\boldsymbol{w}_t - \boldsymbol{w}^U{}_t\| \leq M_1 \frac{r}{n}$ *where* $M_1 = \frac{2}{\mu}c_2$ *is some positive constant that does not depend on* t*.*

To show that the preliminary estimate on the bound on $\|\mathbf{w}^I{}_t - \mathbf{w}_t\|$ used in Theorem 1 and Corollary 1 holds, the proof is provided as below:

Theorem 3 (Bound between iterates on full data and incrementally updated ones)**.** *Consider an iteration* t *indexed with* j_m *for which* $j_m \leq t < j_m + T_0 - 1$*, and suppose that we are at the* x*-th iteration of full gradient updates, so* $j_1 = j_0 + xT_0$*,* $j_m = j_0 + (m - 1 + x)T_0$*. Suppose that we have the bounds* $\|\boldsymbol{H}_t - \boldsymbol{B}_{j_m}\| \leq \xi_{j_1,j_m} = Ad_{j_1,j_m+T_0-1} + \frac{1}{\frac{1}{2}-\frac{r}{n}} AM_1\frac{r}{n}$ *(where we recalled the definition of* ξ*) and* $\xi_{j_1,j_m} \leq \frac{\mu}{2}$ *for all iterations* x*. Then*

$$\|\boldsymbol{w}^I{}_{t+1} - \boldsymbol{w}_{t+1}\| \leq \frac{2rc_2/n}{(1-r/n)\mu - \xi_{j_0,j_0+(m-1)T_0}} \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}.$$

Recall that c_0 *is the Lipshitz constant of the Hessian,* M_1 *and* A *are defined in Theorem 2 and Corollary 1 respectively, which do not depend on* t*,*

For this theorem, note that this inequality depends on the condition $\|\mathbf{H}_t - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m}$ while in Theorem 1, to prove $\|\mathbf{H}_t - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m}$, we need to use the inequality in Theorem 3, i.e. $\|\mathbf{w}^I{}_{t+1} - \mathbf{w}_{t+1}\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$. In what follows, we will show that both inequalities hold for all the iterations t without relying on other conditions.

We can select hyper-parameters T_0, j_0 such that

$$A(1-\eta\mu)^{j_0-m+1} d_{0,(m-1)T_0} + \frac{1}{\frac{1}{2}-\frac{r}{n}} AM_1\frac{r}{n} < \min(\frac{\mu}{2}, (1-\frac{r}{n})\mu - \frac{c_0 M_1 r(n-r)}{2n^2}),$$

e.g. when $m = 2$ and $T_0 = 5$, which is what we used in our experiments. It is enough that

$$j_0 > \max(\frac{\log(\frac{1}{Ad_{0,5}}[\frac{\mu}{2} - \frac{1}{\frac{1}{2}-\frac{r}{n}}AM_1\frac{r}{n})]}{\log(1-\eta\mu)}, \frac{\log(\frac{1}{Ad_{0,5}}[(1-\frac{r}{n})\mu - \frac{1}{\frac{1}{2}-\frac{r}{n}}AM_1\frac{r}{n})]}{\log(1-\eta\mu)}) + m - 1.$$

This holds for small enough r/n:

$$j_0 > \frac{\log(\frac{1}{Ad_{0,5}}[\frac{\mu}{2} - \frac{1}{\frac{1}{2}-\frac{r}{n}}AM_1\frac{r}{n})]}{\log(1-\eta\mu)} + m - 1$$

Then the following two theorems hold.

Theorem 4 (Bound between iterates on full data and incrementally updated ones (all iterations))**.** *For any* $j_m < t < j_m + T_0 - 1$, $\|\boldsymbol{w}^I{}_t - \boldsymbol{w}_t\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}$ *and* $\|\boldsymbol{H}_t - \boldsymbol{B}_{j_m}\| \leq \xi_{j_1,j_m}$.

Then we have the following bound for $\|\mathbf{w}^U{}_t - \mathbf{w}^I{}_t\|$, which is our main result.

Theorem 5 (Convergence rate of DeltaGrad)**.** *For all iterations* t*, the result* $\boldsymbol{w}^I{}_t$ *of DeltaGrad, Algorithm 1, approximates the correct iteration values* $\boldsymbol{w}^U{}_t$ *at the rate*

$$\|\boldsymbol{w}^U{}_t - \boldsymbol{w}^I{}_t\| = o(\frac{r}{n}).$$

So $\|\boldsymbol{w}^U{}_t - \boldsymbol{w}^I{}_t\|$ *is of a lower order than* $\frac{r}{n}$.

This is proved in Section A.2.8.

A.2.5 Proof of Theorem 2

Proof. By subtracting the GD update from equation (1), we have:

$$\begin{aligned}
&\mathbf{w}^U{}_{t+1} - \mathbf{w}_{t+1} = \mathbf{w}^U{}_t - \mathbf{w}_t \\
&- \eta\left(\frac{1}{n-r}\left(\sum_{i=1}^{n}\nabla F_i\left(\mathbf{w}^U{}_t\right) - \sum_{i\in R}\nabla F_i\left(\mathbf{w}^U{}_t\right)\right) - \frac{1}{n}\sum_{i=1}^{n}\nabla F_i\left(\mathbf{w}_t\right)\right)
\end{aligned} \tag{28}$$

in which the right-hand side can be rewritten as:

$$\begin{aligned}
&\mathbf{w}^U{}_t - \mathbf{w}_t - \eta\left(\nabla F(\mathbf{w}^U{}_t) - \nabla F(\mathbf{w}_t)\right) \\
&- \eta\left(\frac{1}{n-r}\left(\sum_{i=1}^{n}\nabla F_i\left(\mathbf{w}^U{}_t\right) - \sum_{i\in R}\nabla F_i\left(\mathbf{w}^U{}_t\right)\right) - \frac{1}{n}\sum_{i=1}^{n}\nabla F_i\left(\mathbf{w}^U{}_t\right)\right) \\
&= \mathbf{w}^U{}_t - \mathbf{w}_t - \eta\left(\nabla F(\mathbf{w}^U{}_t) - \nabla F(\mathbf{w}_t)\right) \\
&- \frac{\eta}{n-r}\left(\frac{r}{n}\sum_{i=1}^{n}\nabla F_i\left(\mathbf{w}^U{}_t\right) - \sum_{i\in R}\nabla F_i\left(\mathbf{w}^U{}_t\right)\right) \\
&= \mathbf{w}^U{}_t - \mathbf{w}_t - \eta\left(\nabla F(\mathbf{w}^U{}_t) - \nabla F(\mathbf{w}_t)\right) + \delta_t.
\end{aligned}$$

Then by applying Cauchy mean value theorem, the triangle inequality, Cauchy schwarz inequality and Lemma 7 respectively, we have:

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}^U{}_{t+1}\| \\
&\leq \|\mathbf{w}_t - \mathbf{w}^U{}_t - \eta(\int_0^1 \mathbf{H}\left(\mathbf{w}_t + x\left(\mathbf{w}^U{}_t - \mathbf{w}_t\right)\right)dx)\left(\mathbf{w}_t - \mathbf{w}^U{}_t\right)\| + \|\delta_t\| \\
&\leq \|\mathbf{I} - \eta\int_0^1 \mathbf{H}\left(\mathbf{w}_t + x\left(\mathbf{w}^U{}_t - \mathbf{w}_t\right)\right)dx\|\|\mathbf{w}_t - \mathbf{w}^U{}_t\| + \frac{2c_2 r\eta}{n}
\end{aligned}$$

Then by applying the triangle inequality over integrals and Lemma 2, the formula can be further bounded as:

$$\begin{aligned}
&\leq \|\int_0^1 (\mathbf{I} - \eta \mathbf{H}\left(\mathbf{w}_t + x\left({\mathbf{w}^U}_t - \mathbf{w}_t\right)\right) dx)\| \|\mathbf{w}_t - {\mathbf{w}^U}_t\| + \frac{2c_2 r\eta}{n} \\
&\leq (1 - \eta\mu)\|\mathbf{w}_t - {\mathbf{w}^U}_t\| + \frac{2c_2 r\eta}{n}
\end{aligned}$$

Then by applying this formula iteratively, we get:

$$\|\mathbf{w}_{t+1} - {\mathbf{w}^U}_{t+1}\| \leq \frac{1}{\eta\mu}\frac{2c_2 r\eta}{n} = \frac{2c_2}{\mu}\frac{r}{n} := M_1\frac{r}{n}$$

□

A.2.6 Proof of Theorem 3

Proof. The updates for the iterations $j_m \leq t \leq j_m + T_0 - 1$ follow the Quasi-Hessian update. We proceed in a similar way as before, by expanding the recursion as below:

$$\begin{aligned}
&\|{\mathbf{w}^I}_{t+1} - \mathbf{w}_{t+1}\| \\
&= \|{\mathbf{w}^I}_t - (\mathbf{w}_t - \eta\nabla F\left(\mathbf{w}_t\right)) \\
&- \frac{\eta}{n-r}(n\left[\mathbf{B}_{j_m}\left({\mathbf{w}^I}_t - \mathbf{w}_t\right) + \nabla F\left(\mathbf{w}_t\right)\right] - \sum_{i\in R}\nabla F_i\left({\mathbf{w}^I}_t\right))\| \\
&= \|(\mathbf{I} - \eta\frac{n}{n-r}\mathbf{B}_{j_m})({\mathbf{w}^I}_t - \mathbf{w}_t) - \frac{r\eta}{n-r}\nabla F(\mathbf{w}_t) + \frac{\eta}{n-r}\sum_{i\in R}\nabla F_i({\mathbf{w}^I}_t)\|
\end{aligned} \tag{29}$$

By rearranging the formula above and using the triangle inequality, we get:

$$\begin{aligned}
&= \|(\mathbf{I} - \eta\frac{n}{n-r}\mathbf{B}_{j_m})({\mathbf{w}^I}_t - \mathbf{w}_t) - \frac{r\eta}{n-r}\nabla F(\mathbf{w}_t) \\
&+ \frac{\eta}{n-r}\sum_{i\in R}(\mathbf{H}_{t,i} \times ({\mathbf{w}^I}_t - \mathbf{w}_t) + \nabla F_i(\mathbf{w}_t))\| \\
&\leq \|(\mathbf{I} - \eta\frac{n}{n-r}\mathbf{B}_{j_m})({\mathbf{w}^I}_t - \mathbf{w}_t) + \frac{\eta}{n-r}\sum_{i\in R}\mathbf{H}_{t,i} \times ({\mathbf{w}^I}_t - \mathbf{w}_t)\| \\
&+ \|\frac{r\eta}{n-r}\nabla F(\mathbf{w}_t)\| + \|\frac{\eta}{n-r}\sum_{i\in R}\nabla F_i(\mathbf{w}_t)\|
\end{aligned} \tag{30}$$

in which we use $\mathbf{H}_{t,i}$ to denote $\int_0^1 \mathbf{H}_i(\mathbf{w}_t + x({\mathbf{w}^I}_t - \mathbf{w}_t))dx$ (recall that $\mathbf{H}_i$ represents the Hessian matrix evaluated at the i_{th} sample). Then the terms in the first absolute value are rewritten as:

$$\begin{aligned}
&[\mathbf{I} - \eta\frac{n}{n-r}(\mathbf{B}_{j_m} - \mathbf{H}_t + \mathbf{H}_t)]({\mathbf{w}^I}_t - \mathbf{w}_t) + \frac{\eta}{n-r}\sum_{i\in R}\mathbf{H}_{t,i} \times ({\mathbf{w}^I}_t - \mathbf{w}_t) \\
&= [\mathbf{I} - \eta\frac{n}{n-r}(\mathbf{B}_{j_m} - \mathbf{H}_t)]({\mathbf{w}^I}_t - \mathbf{w}_t) - \frac{\eta}{n-r}\sum_{i\notin R}\mathbf{H}_{t,i} \times ({\mathbf{w}^I}_t - \mathbf{w}_t)
\end{aligned}$$

which uses the fact that $\mathbf{H}_t = \sum_{i=1}^n \mathbf{H}_{t,i} = \sum_{i\notin R}\mathbf{H}_{t,i} + \sum_{i\in R}\mathbf{H}_{t,i}$. Then Formula (30) can be further bounded as:

$$\begin{aligned}
&\leq \|[\mathbf{I} - \frac{\eta}{n-r}\sum_{i\notin R}\mathbf{H}_{t,i}]({\mathbf{w}^I}_t - \mathbf{w}_t)\| + \frac{n\eta}{n-r}\|(\mathbf{B}_{j_m} - \mathbf{H}_t)({\mathbf{w}^I}_t - \mathbf{w}_t)\| \\
&+ \|\frac{r\eta}{n-r}\nabla F(\mathbf{w}_t)\| + \|\frac{\eta}{n-r}\sum_{i\in R}\nabla F_i(\mathbf{w}_t)\| \\
&\leq (1 - \eta\mu + \eta\frac{n}{n-r}\xi_{j_1,j_m})\|{\mathbf{w}^I}_t - \mathbf{w}_t\| + \frac{r\eta c_2}{n-r} + \frac{\eta r c_2}{n-r}
\end{aligned} \tag{31}$$

Then according to Lemma 8, $d_{j_1,j_m+T_0-1} = d_{j_0+xT_0,j_0+(x+m)T_0-1}$ decreases with increasing x, and thus $\xi_{j_1,j_m} = Ad_{j_1,j_m+T_0-1} + \frac{1}{\frac{1}{2}-\frac{r}{n}} AM_1 \frac{r}{n}$ is also decreasing with increasing x. So the formula above can be further bounded as:

$$\leq (1 - \eta\mu + \eta \frac{n}{n-r} \xi_{j_0,j_0+(m-1)T_0}) \|\mathbf{w}^I{}_t - \mathbf{w}_t\| + \frac{2r\eta c_2}{n-r}$$

This shows a recurrent inequality for $\|\mathbf{w}^I{}_t - \mathbf{w}_t\|$. Next, notice that the conditions for deriving the above inequality hold for all $j_m \leq t \leq j_m + T_0 - 1$.

Then, when we reach $t = j_m$, we have an iteration where the gradient is computed exactly. For these iterations we have $\mathbf{w}^I{}_{t+1} = \mathbf{w}^I{}_t - \frac{\eta}{n-r} \sum_{i \notin R} \nabla F(\mathbf{w}^I{}_t)$ as well as $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla F(\mathbf{w}_t)$. Using the same argument as in the bound for $\mathbf{w}_t - \mathbf{w}^U{}_t$ we can get:

$$\|\mathbf{w}_{t+1} - \mathbf{w}^I{}_{t+1}\| \leq [1 - \eta\mu] \|\mathbf{w}_t - \mathbf{w}^I{}_t\| + \frac{2c_2 r\eta}{n}.$$

Therefore, we effectively have $\xi = 0$ for these iterations. We then continue with $t \leftarrow t - 1$, and use the appropriate bound among the two derived above. This recursive process works until we reach $t = 1$.

As long as $\xi_{j_0,j_0+(m-1)T_0} \leq \frac{\mu}{2}$, $-\eta\mu + \eta \frac{n}{n-r} \xi_{j_0,j_0+(m-1)T_0} < -\eta\mu + \eta \frac{n}{n-r} \frac{\mu}{2} < 0$. Then we get the following inequality:

$$\begin{aligned} \|\mathbf{w}^I{}_t - \mathbf{w}_t\| &\leq \frac{2\frac{\eta r c_2}{n-r}}{\eta\mu - \eta \frac{n}{n-r} \xi_{j_0,j_0+(m-1)T_0}} \\ &= \frac{2rc_2/n}{(1-r/n)\mu - \xi_{j_0,j_0+(m-1)T_0}} \end{aligned}$$

As long as $\xi_{j_0,j_0+(m-1)T_0} \leq \frac{\mu}{2}$, then

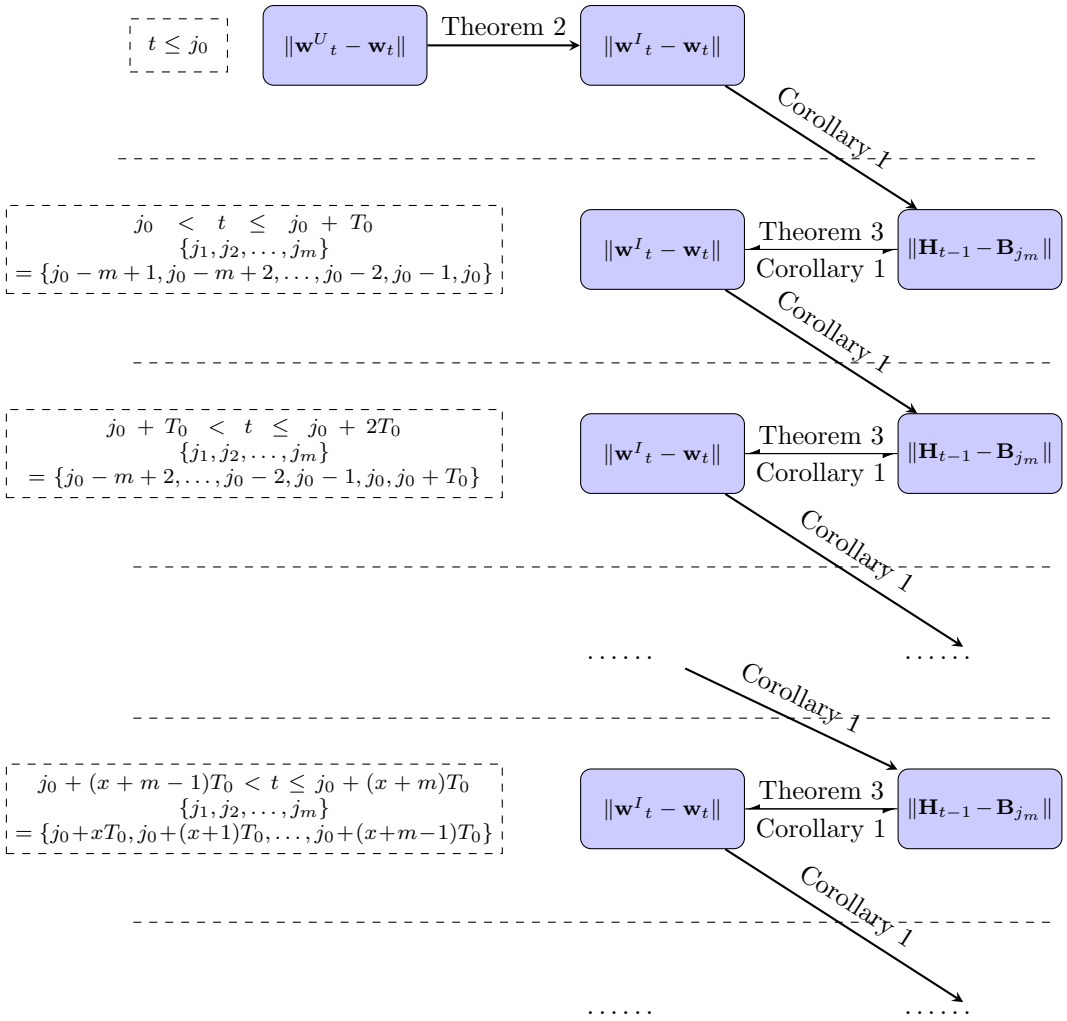
$$\|\mathbf{w}^I{}_t - \mathbf{w}_t\| \leq \frac{2rc_2/n}{(1-r/n)\mu - \xi_{j_0,j_0+(m-1)T_0}} \leq \frac{2rc_2/n}{(1-r/n)\mu - \frac{\mu}{2}} = \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}.$$

The last step uses the fact that $M_1 = \frac{2c_2}{\mu}$.

□

A.2.7 Proof of Theorem 4

Architecture of the proof. To visualize the recursive proof process, we draw a picture as:



Proof. First of all, in terms of the bound on ξ_{j_1,j_m} which is required in Theorem 3, i.e. $\xi_{j_1,j_m} \le \frac{\mu}{2}$, we do the analysis below to show that we can adjust the value of j_0 and T_0 such that it can hold for all t. When $j_1 \ge j_0$, i.e. $j_1 = j_0 + xT_0$, then

$$\{j_2, j_3, \ldots, j_m\} = \{j_0 + (x+1)T_0, j_0 + (x+2)T_0, \ldots, j_0 + (x+m-1)T_0\},$$

thus $\xi_{j_1,j_m} = \xi_{j_0+xT_0,j_0+(x+m-1)T_0} = Ad_{j_0+xT_0,j_0+(x+m)T_0-1} + \frac{1}{\frac{1}{2}-\frac{r}{n}} AM_1 \frac{r}{n}$. Here $d_{j_0+xT_0,j_0+(x+m)T_0-1}$ decreases with x, and so does $\xi_{j_1,j_m} = \xi_{j_0+xT_0,j_0+(x+m-1)T_0}$. So the following inequality holds:

$$d_{j_0+xT_0,j_0+(x+m)T_0-1} \le d_{j_0,j_0+mT_0-1} \le (1-\mu\eta)^{j_0} d_{0,mT_0-1}.$$

When $j_1 < j_0$, there are only m different choices for $\{j_1, j_2, \ldots, j_m\}$, in which the smallest j_1 used for approximation is $j_0 - m + 1$. Then, the following inequality holds:

$$d_{j_1,j_m} \le (1-\eta\mu)^{j_1} d_{0,j_m-j_1} \le (1-\eta\mu)^{j_0-m+1} d_{0,j_m-j_1}.$$

For those $j_1, j_2, \ldots, j_m$, we have $j_m - j_1 \le (m-1)T_0$ and thus

$$d_{j_1,j_m} \le (1-\eta\mu)^{j_1} d_{0,j_m-j_1} \le (1-\eta\mu)^{j_0-m+1} d_{0,j_m-j_1} \le (1-\eta\mu)^{j_0-m+1} d_{0,(m-1)T_0}.$$

So ξ_{j_1,j_m} is bounded by $A(1-\eta\mu)^{j_0-m+1} d_{0,(m-1)T_0} + \frac{1}{\frac{1}{2}-\frac{r}{n}} AM_1 \frac{r}{n}$. To make sure $\xi_{j_1,j_m} \le \frac{\mu}{2}$, we can adjust j_0, m, T_0 to make $A(1-\eta\mu)^{j_0-m+1} d_{0,(m-1)T_0} + \frac{1}{\frac{1}{2}-\frac{r}{n}} AM_1 \frac{r}{n}$ smaller than $\frac{\mu}{2}$.

Then when $t \leq j_0$, the gradient is evaluated explicitly, which means that ${\mathbf{w}^U}_t = {\mathbf{w}^I}_t$, so the bound clearly holds, i.e., from Theorem 2, we have $\|{\mathbf{w}^U}_t - \mathbf{w}_t\| \leq \frac{M_1 r}{n}$ and thus $\|{\mathbf{w}^I}_t - \mathbf{w}_t\| = \|{\mathbf{w}^U}_t - \mathbf{w}_t\| \leq \frac{M_1 r}{n} \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$.

When $j_0 < t < j_0 + T_0$, in order to compute ${\mathbf{w}^I}_t$, we need to use the history information $\{\Delta w_{j_1}, \Delta w_{j_2}, \ldots, \Delta w_{j_m}\}$, $\{\Delta g_{j_1}, \Delta g_{j_2}, \ldots, \Delta g_{j_m}\}$ and the corresponding quasi-Hessian matrices $\{\mathbf{B}_{j_1}, \mathbf{B}_{j_2}, \ldots, \mathbf{B}_{j_m}\}$ where $\{j_1, j_2, \ldots, j_m\} = \{j_0 - m + 1, j_0 - m + 2, \ldots, j_0\}$ (we suppose $m < j_0$, which is a natural assumption). Since $\|{\mathbf{w}^I}_t - \mathbf{w}_t\| \leq \frac{M_1 r}{n}$ for any $t \leq j_0$, the conditions of Corollary 1 (used here with the $j_1, \ldots, j_m$ described above) hold up to j_0, so when $t = j_0 + 1$, $\|\mathbf{H}_{t-1} - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m}$ where

$$\xi_{j_1,j_m} = \xi_{j_0-m+1,j_0} = Ad_{j_1,j_m+T_0-1} + AM_1\frac{r}{n} = Ad_{j_0-m+1,j_0+T_0} + AM_1\frac{r}{n}.$$

Plus, according to Theorem 3, $\|{\mathbf{w}^I}_t - \mathbf{w}_t\| \leq \frac{2rc_2/n}{(1-r/n)\mu - \xi_{j_1,j_m}} = \frac{2rc_2/n}{(1-r/n)\mu - \xi_{j_0-m+1,j_0}}$. When $\xi_{j_0-m+1,j_0} \leq \frac{\mu}{2}$, then

$$\|{\mathbf{w}^I}_t - \mathbf{w}_t\| \leq \frac{2rc_2/n}{(1-r/n)\mu - \xi_{j_1,j_m}} = \frac{2rc_2/n}{(1-r/n)\mu - \frac{\mu}{2}} = \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}.$$

So the bound on $\|{\mathbf{w}^I}_t - \mathbf{w}_t\|$ holds for all $t \leq j_0 + 1$. Then according to the conditions of Corollary 1, when $t = j_0 + 2$, $\|\mathbf{H}_{t-1} - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m}$ holds. This can proceed recursively until $t = j_0 + T_0$, in which the gradients are explicitly evaluated according to Theorem 3, i.e.:

$$\|{\mathbf{w}^I}_{j_0+T_0} - \mathbf{w}_{j_0+T_0}\| \leq \frac{2rc_2/n}{(1-r/n)\mu - \xi_{j_1,j_m}} = \frac{2rc_2/n}{(1-r/n)\mu - \frac{\mu}{2}} = \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}.$$

Next when $j_0 + T_0 < t < j_0 + 2T_0$, j_m is updated as $j_0 + T_0$ while $\{j_1, j_2, \ldots, j_{m-1}\}$ is updated as $\{j_0 - m + 2, j_0 - m + 3, \ldots, j_0\}$ and we know that $\|{\mathbf{w}^I}_{j_k} - \mathbf{w}_{j_k}\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$. So based on Corollary 1, the following inequality holds:

$$\|\mathbf{H}_t - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m} = \xi_{j_0-m+2,j_0+T_0} = Ad_{j_1,j_m+T_0-1} + AM_1\frac{r}{n}$$

$$= Ad_{j_0-m+2,j_0+2T_0} + AM_1\frac{r}{n}$$

This process can proceed recursively.

When $j_0 + xT_0 < t < j_0 + (x+1)T_0$, we know that:

$$\|\mathbf{H}_t - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m} = Ad_{j_1,j_m+T_0-1} + AM_1\frac{r}{n}.$$

Then based on Theorem 3, $\|{\mathbf{w}^I}_t - \mathbf{w}_t\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$. Then at iteration $j_0 + (x+1)T_0$, we update $j_1, j_2, \ldots, j_m$ as: $j_m \leftarrow j_0 + (x+1)T_0$ and $j_{i-1} \leftarrow j_i$ $(i = 2, 3, \ldots, m)$ and thus

$$\|{\mathbf{w}^I}_{j_k} - \mathbf{w}_{j_k}\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$$

still holds for all $k = 1, 2, \ldots, m$.

So when $j_0 + (x+1)T_0 < t < j_0 + (x+2)T_0$, Corollary 1 and Theorem 3 are applied alternatively. Then the following two inequalities hold for all iterations t satisfying $j_0 + (x+1)T_0 < t < j_0 + (x+2)T_0$:

$$\|\mathbf{H}_t - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m} = Ad_{j_1,j_m+T_0-1} + AM_1\frac{r}{n},$$

$$\|{\mathbf{w}^I}_{j_k} - \mathbf{w}_{j_k}\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}.$$

So in the end, we know that:

$$\|{\mathbf{w}^I}_t - \mathbf{w}_t\| \leq \frac{1}{\frac{1}{2}-\frac{r}{n}} M_1 \frac{r}{n}$$

and

$$\|\mathbf{H}_t - \mathbf{B}_{j_m}\| \leq \xi_{j_1,j_m}$$

hold for all t.

□

A.2.8 Proof of Theorem 5

Proof. The proof is by induction.

When $t \leq j_0$, the gradient is evaluated explicitly, which means that ${\mathbf{w}^U}_t = {\mathbf{w}^I}_t$, so the bound clearly holds.

From iteration T_0 to iteration t, the difference between ${\mathbf{w}^I}_t$ and ${\mathbf{w}^U}_t$ can be bounded as follows. In these equations, we use the definition of the update formula ${\mathbf{w}^I}_{t+1} = {\mathbf{w}^I}_t - \frac{\eta}{n-r}[n(\mathbf{B}_{j_m}({\mathbf{w}^I}_t - \mathbf{w}_t)) - \sum_{i \in R} \nabla F({\mathbf{w}^I}_t)]$. By rearranging terms appropriately, we get:

$$\begin{aligned} \|{\mathbf{w}^I}_{t+1} - {\mathbf{w}^U}_{t+1}\| = \|{\mathbf{w}^I}_t - {\mathbf{w}^U}_t - \frac{n\eta}{n-r}\left[\mathbf{B}_{j_m}({\mathbf{w}^I}_t - \mathbf{w}_t) + \nabla F(\mathbf{w}_t)\right] \\ + \frac{\eta}{n-r}\sum_{i \in R} \nabla F_i({\mathbf{w}^I}_t) + \frac{\eta}{n-r}\sum_{i \notin R} \nabla F_i({\mathbf{w}^U}_t)\| \end{aligned} \tag{32}$$

Then by bringing in $\mathbf{H}_t$ into the expression above, it is rewritten as:

$$\begin{aligned} &= \|{\mathbf{w}^I}_t - {\mathbf{w}^U}_t - \frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m} - \mathbf{H}_t\right)\left({\mathbf{w}^I}_t - \mathbf{w}_t\right) + \mathbf{H}_t \times \left({\mathbf{w}^I}_t - \mathbf{w}_t\right) + \nabla F(\mathbf{w}_t)\right] \\ &+ \frac{\eta}{n-r}\sum_{i \in R}(\nabla F_i({\mathbf{w}^I}_t) - \nabla F_i(\mathbf{w}_t) + \nabla F_i(\mathbf{w}_t)) + \frac{\eta}{n-r}\sum_{i \notin R} \nabla F_i({\mathbf{w}^U}_t)\| \end{aligned} \tag{33}$$

In the formula above, we will try to make sure there is no confusion between $\mathbf{H}_t(\mathbf{w})$ (Hessian as a function evaluated at $\mathbf{w}$) and $\mathbf{H}_t \times (\mathbf{w})$ (Hessian times a vector). Then by applying the Cauchy mean value theorem over each individual $\nabla F_i({\mathbf{w}^I}_t) - \nabla F_i(\mathbf{w}_t)$ and by denoting the corresponding Hessian matrix as $\mathbf{H}_{t,i}$ (note that $\sum_{i=1}^{n} \mathbf{H}_{t,i} = n\mathbf{H}_t$), the expression becomes:

$$\begin{aligned} &= \|{\mathbf{w}^I}_t - {\mathbf{w}^U}_t - \frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m} - \mathbf{H}_t\right)\left({\mathbf{w}^I}_t - \mathbf{w}_t\right) + \mathbf{H}_t \times \left({\mathbf{w}^I}_t - \mathbf{w}_t\right) + \nabla F(\mathbf{w}_t)\right] \\ &+ \frac{\eta}{n-r}\sum_{i \in R}(\mathbf{H}_{t,i} \times \left({\mathbf{w}^I}_t - \mathbf{w}_t\right) + \nabla F_i(\mathbf{w}_t)) + \frac{\eta}{n-r}\sum_{i \notin R} \nabla F_i({\mathbf{w}^U}_t)\| \end{aligned}$$

Then by using the fact that $\sum_{i \in R} \nabla F_i(\mathbf{w}_t) + \sum_{i \notin R} \nabla F_i(\mathbf{w}_t) = n\nabla F(\mathbf{w}_t)$ and $\sum_{i \in R} \mathbf{H}_{t,i} + \sum_{i \notin R} \mathbf{H}_{t,i} = n\mathbf{H}_t$, the expression can be rearranged as:

$$\begin{aligned} &= \|{\mathbf{w}^I}_t - {\mathbf{w}^U}_t - \frac{\eta}{n-r}\sum_{i \notin R} \mathbf{H}_{t,i} \times \left({\mathbf{w}^I}_t - \mathbf{w}_t\right) - \frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m} - \mathbf{H}_t\right)\left({\mathbf{w}^I}_t - \mathbf{w}_t\right)\right] \\ &- \frac{\eta}{n-r}\sum_{i \notin R} \nabla F_i(\mathbf{w}_t) + \frac{\eta}{n-r}\sum_{i \notin R} \nabla F_i({\mathbf{w}^U}_t)\| \end{aligned}$$

in which $\frac{\eta}{n-r}\sum_{i \in R} \nabla F_i(\mathbf{w}_t)$ is canceled out. Then by adding and subtracting ${\mathbf{w}^U}_t$ in the first part, we get:

$$\begin{aligned} &= \|{\mathbf{w}^I}_t - {\mathbf{w}^U}_t - \frac{\eta}{n-r}\sum_{i \notin R} \mathbf{H}_{t,i} \times \left({\mathbf{w}^I}_t - {\mathbf{w}^U}_t\right) - \frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m} - \mathbf{H}_t\right)\left({\mathbf{w}^I}_t - {\mathbf{w}^U}_t\right)\right] \\ &- \frac{\eta}{n-r}\sum_{i \notin R} \mathbf{H}_{t,i} \times \left({\mathbf{w}^U}_t - \mathbf{w}_t\right) - \frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m} - \mathbf{H}_t\right)\left({\mathbf{w}^U}_t - \mathbf{w}_t\right)\right] \\ &- \frac{\eta}{n-r}\sum_{i \notin R} \nabla F_i(\mathbf{w}_t) + \frac{\eta}{n-r}\sum_{i \notin R} \nabla F_i({\mathbf{w}^U}_t)\| \end{aligned}$$

We apply Cauchy mean value theorem over $-\frac{\eta}{n-r}\sum_{i\notin R}\nabla F_i(\mathbf{w}_t)+\frac{\eta}{n-r}\sum_{i\notin R}\nabla F_i(\mathbf{w}^U{}_t)$, i.e.:

$$
\begin{aligned}
&-\frac{\eta}{n-r}\sum_{i\notin R}\nabla F_i(\mathbf{w}_t)+\frac{\eta}{n-r}\sum_{i\notin R}\nabla F_i(\mathbf{w}^U{}_t)\\
&=\frac{\eta}{n-r}[\sum_{i\notin R}\int_0^1\mathbf{H}_i(\mathbf{w}_t+x(\mathbf{w}^U{}_t-\mathbf{w}_t))dx](\mathbf{w}^U{}_t-\mathbf{w}_t).
\end{aligned}
$$

In addition, note that $\mathbf{H}_{t,i}=\int_0^1\mathbf{H}_i(\mathbf{w}_t+x(\mathbf{w}^I{}_t-\mathbf{w}_t))dx$. So the formula above becomes:

$$
\begin{aligned}
&=\|\mathbf{w}^I{}_t-\mathbf{w}^U{}_t-\frac{\eta}{n-r}\sum_{i\notin R}\mathbf{H}_{t,i}\times(\mathbf{w}^I{}_t-\mathbf{w}^U{}_t)-\frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m}-\mathbf{H}_t\right)(\mathbf{w}^I{}_t-\mathbf{w}^U{}_t)\right]\\
&-\frac{\eta}{n-r}\sum_{i\notin R}(\int_0^1\mathbf{H}_i(\mathbf{w}_t+x(\mathbf{w}^I{}_t-\mathbf{w}_t))dx)(\mathbf{w}^U{}_t-\mathbf{w}_t)-\frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m}-\mathbf{H}_t\right)(\mathbf{w}^U{}_t-\mathbf{w}_t)\right]\\
&+\frac{\eta}{n-r}\sum_{i\notin R}(\int_0^1\mathbf{H}_i(\mathbf{w}_t+x(\mathbf{w}^U{}_t-\mathbf{w}_t))dx)(\mathbf{w}^U{}_t-\mathbf{w}_t)\|.
\end{aligned}
$$

Then by applying the triangle inequality and rearranging the expression appropriately, the expression can be bounded as:

$$
\begin{aligned}
&\leq\|(\mathbf{I}-\frac{\eta}{n-r}\sum_{i\notin R}\mathbf{H}_{t,i})(\mathbf{w}^I{}_t-\mathbf{w}^U{}_t)\|+\|\frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m}-\mathbf{H}_t\right)(\mathbf{w}^I{}_t-\mathbf{w}^U{}_t)\right]\|\\
&+\|\frac{\eta}{n-r}[\sum_{i\notin R}\int_0^1\mathbf{H}_i(\mathbf{w}_t+x(\mathbf{w}^U{}_t-\mathbf{w}_t))dx-\int_0^1\mathbf{H}_i(\mathbf{w}_t+x(\mathbf{w}^I{}_t-\mathbf{w}_t))dx](\mathbf{w}^U{}_t-\mathbf{w}_t)\|\\
&+\|\frac{n\eta}{n-r}\left[\left(\mathbf{B}_{j_m}-\mathbf{H}_t\right)(\mathbf{w}^U{}_t-\mathbf{w}_t)\right]\|,
\end{aligned}
$$

in which the first term is the main contraction component which always appears in the analyses of gradient descent type algorithms. The remaining terms are error terms due to the various sources of error: using a quasi-Hessian, not having a quadratic objective (implicitly assumed by the local models at each step), using the iterate $\mathbf{w}^I$ for our update instead of the correct $\mathbf{w}^U$.

Then by using the following facts:

1. $\|\mathbf{I}-\eta\mathbf{H}_{t,i}\|\leq 1-\eta\mu$;
2. from Theorem 4 on the approximation accuracy of the quasi-Hessian to mean Hessian, we have the error bound $\|\mathbf{H}_t-\mathbf{B}_{j_m}\|\leq\xi_{j_1,j_m}$;
3. we can bound the difference of integrated Hessians using the strategy from equation (20);
4. from Theorem 2, we have the error bound $\|\mathbf{w}^U{}_t-\mathbf{w}_t\|\leq M_1\frac{r}{n}$ (and this requires no additional assumptions),

the expression above can be bounded as follows:

$$
\begin{aligned}
&\leq(1-\eta\mu+\frac{n\eta}{n-r}\xi_{j_1,j_m})\|\mathbf{w}^I{}_t-\mathbf{w}^U{}_t\|+\frac{\eta c_0}{2}\|\mathbf{w}^U{}_t-\mathbf{w}^I{}_t\|\|\mathbf{w}^U{}_t-\mathbf{w}_t\|\\
&+\frac{n\eta}{n-r}\xi_{j_1,j_m}\|\mathbf{w}^U{}_t-\mathbf{w}_t\|\\
&\leq(1-\eta\mu+\frac{n\eta}{n-r}\xi_{j_1,j_m}+\frac{c_0M_1r\eta}{2n})\|\mathbf{w}^I{}_t-\mathbf{w}^U{}_t\|+\frac{M_1r\eta}{n-r}\xi_{j_1,j_m}
\end{aligned}
\tag{34}
$$

Recall from Corollary 1 that $\xi_{j_1,j_m} = \xi_{j_0+xT_0,j_0+(m+x-1)T_0} = Ad_{j_0+xT_0,j_0+(m+x)T_0-1} + A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}$ decreases with the increasing x. So the formula above can be bounded as:

$$\leq (1-\eta\mu+\frac{n\eta}{n-r}\xi_{j_0,j_0+(m-1)T_0}+\frac{c_0M_1r\eta}{2n})\|\mathbf{w}^I{}_t-\mathbf{w}^U{}_t\|+\frac{M_1r\eta}{n-r}\xi_{j_1,j_m}. \tag{35}$$

Also by plugging the formula for ξ into the formula above and using Lemma 8 (contraction of GD updates), we get:

$$\begin{aligned}
&\leq (1-\eta\mu+\frac{n\eta}{n-r}\xi_{j_0,j_0+(m-1)T_0}+\frac{c_0M_1r\eta}{2n})\|\mathbf{w}^I{}_t-\mathbf{w}^U{}_t\| \\
&+\frac{M_1r\eta}{n-r}(Ad_{j_0+xT_0,j_0+(m+x)T_0-1}+A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}) \\
&\leq (1-\eta\mu+\frac{n\eta}{n-r}\xi_{j_0,j_0+(m-1)T_0}+\frac{c_0M_1r\eta}{2n})\|\mathbf{w}^I{}_t-\mathbf{w}^U{}_t\| \\
&+\frac{M_1r\eta}{n-r}(A(1-\eta\mu)^{j_0+xT_0}d_{0,mT_0-1}+A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n})
\end{aligned} \tag{36}$$

Now, we will argue that it is pssible to choose hyperparameters such that $\xi_{j_0,j_0+(m-1)T_0} \leq (1-\frac{r}{n})\mu - \frac{c_0M_1r(n-r)}{2n^2}$. Then $1-\eta\mu+\frac{n\eta}{n-r}\xi_{j_0,j_0+(m-1)T_0}+\frac{c_0M_1r\eta}{2n}$ is a constant for all t and smaller than 1. By denoting $\mu-\frac{n}{n-r}\xi_{j_0,j_0+(m-1)T_0}-\frac{c_0M_1r}{2n}$ as C, the formula above can be written as:

$$=(1-\eta C)\|\mathbf{w}^I{}_t-\mathbf{w}^U{}_t\|+\frac{M_1r\eta}{n-r}(A(1-\eta\mu)^{j_0+xT_0}d_{0,mT_0-1}+A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}).$$

This can be used recursively until iteration $j_m = j_0+(x+m)T_0-1$, i.e.:

$$\begin{aligned}
&\leq (1-\eta C)^{t-(j_0+(x+m-1)T_0)-1}\|\mathbf{w}^I{}_{j_0+(x+m-1)T_0+1}-\mathbf{w}^U{}_{j_0+(x+m-1)T_0+1}\| \\
&+\frac{1-(1-\eta C)^{t-(j_0+(x+m-1)T_0)}}{\eta C}\frac{M_1r\eta}{n-r}(A(1-\eta\mu)^{j_0+xT_0}d_{0,mT_0-1}+A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}) \\
&\leq (1-\eta C)^{t-(j_0+(x+m-1)T_0)-1}\|\mathbf{w}^I{}_{j_0+(x+m-1)T_0+1}-\mathbf{w}^U{}_{j_0+(x+m-1)T_0+1}\| \\
&+\frac{M_1r}{C(n-r)}(A(1-\eta\mu)^{j_0+xT_0}d_{0,mT_0-1}+A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n})
\end{aligned}$$

We can set $t = j_0+(y+m)T_0$ and for any $y = 1,2,\ldots,x-1$, the formula above can be rewritten as:

$$\begin{aligned}
&\|\mathbf{w}^I{}_{j_0+(y+m)T_0}-\mathbf{w}^U{}_{j_0+(y+m)T_0}\| \\
&\leq (1-\eta C)^{T_0-1}\|\mathbf{w}^I{}_{j_0+(y+m-1)T_0+1}-\mathbf{w}^U{}_{j_0+(y+m-1)T_0+1}\| \\
&+\frac{M_1r}{C(n-r)}(A(1-\eta\mu)^{j_0+yT_0}d_{0,mT_0-1}+A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n})
\end{aligned}$$

Then at the iteration $t = j_0+(y+m-1)T_0$, the gradient is explicitly evaluated, which means that:

$$\|\mathbf{w}^I{}_{j_0+(y+m-1)T_0+1}-\mathbf{w}^U{}_{j_0+(y+m-1)T_0+1}\| \leq (1-\eta\mu)\|\mathbf{w}^I{}_{j_0+(y+m-1)T_0}-\mathbf{w}^U{}_{j_0+(y+m-1)T_0}\|.$$

Since $C = \mu-\frac{n}{n-r}\xi_{j_0,j_0+(m-1)T_0}-\frac{c_0M_1r}{2n}$, then $1-\eta\mu < 1-\eta C$ and thus

$$\|\mathbf{w}^I{}_{j_0+(y+m-1)T_0+1}-\mathbf{w}^U{}_{j_0+(y+m-1)T_0+1}\| \leq (1-\eta C)\|\mathbf{w}^I{}_{j_0+(y+m-1)T_0}-\mathbf{w}^U{}_{j_0+(y+m-1)T_0}\|,$$

which can be plugged into the formula above:

$$\begin{aligned}
&\|\mathbf{w}^I{}_{j_0+(y+m)T_0}-\mathbf{w}^U{}_{j_0+(y+m)T_0}\| \\
&\leq (1-\eta C)^{T_0}\|\mathbf{w}^I{}_{j_0+(y+m-1)T_0}-\mathbf{w}^U{}_{j_0+(y+m-1)T_0}\| \\
&+\frac{M_1r}{C(n-r)}(A(1-\eta\mu)^{j_0+yT_0}d_{0,mT_0-1}+A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}).
\end{aligned}$$

This can be used recursively over $y = x-1, x-2, \ldots, 2, 1$:

$$\begin{aligned}
&\|\mathbf{w}^I{}_{j_0+(y+m)T_0} - \mathbf{w}^U{}_{j_0+(y+m)T_0}\| \\
&\leq (1-\eta C)^{yT_0}\|\mathbf{w}^I{}_{j_0+mT_0} - \mathbf{w}^U{}_{j_0+mT_0}\| \\
&+ \sum_{p=1}^{y}(1-\eta C)^{(y-p)T_0}\frac{M_1 r}{C(n-r)}(A(1-\eta\mu)^{j_0+pT_0}d_{0,mT_0-1} + A\frac{1}{\frac{1}{2}-\frac{r}{n}}M_1\frac{r}{n}) \\
&= (1-\eta C)^{yT_0}\|\mathbf{w}^I{}_{j_0+mT_0} - \mathbf{w}^U{}_{j_0+mT_0}\| \\
&+ \sum_{p=1}^{y}(1-\eta C)^{(y-p)T_0}\frac{M_1 r}{C(n-r)}(A(1-\eta\mu)^{j_0+pT_0}d_{0,mT_0-1}) \\
&+ \sum_{p=1}^{y}(1-\eta C)^{(y-p)T_0}\frac{AM_1^2 r^2}{C(n-r)(n/2-r)},
\end{aligned} \tag{37}$$

in which

$$\sum_{p=1}^{y}(1-\eta C)^{(y-p)T_0}\frac{M_1 r}{C(n-r)}(A(1-\eta\mu)^{j_0+pT_0}d_{0,mT_0-1})$$

$$= \frac{AM_1 r\eta}{C(n-r)}(1-\eta C)^{yT_0}(1-\eta\mu)^{j_0}d_{0,mT_0-1}\sum_{p=1}^{y}(1-\eta C)^{-pT_0}(1-\eta\mu)^{pT_0}.$$

Recall that since $1-\eta C > 1-\eta\mu$, then the formula above can be bounded as:

$$\sum_{p=1}^{y}(1-\eta C)^{(y-p)T_0}\frac{M_1 r}{C(n-r)}(A(1-\eta\mu)^{j_0+pT_0}d_{0,mT_0-1})$$

$$\leq \frac{AM_1 r\eta}{C(n-r)}(1-\eta C)^{yT_0}(1-\eta\mu)^{j_0}d_{0,mT_0-1}\frac{1}{1-(\frac{1-\eta\mu}{1-\eta C})^{T_0}}.$$

Also $\sum_{p=1}^{y}(1-\eta C)^{(y-p)T_0}\frac{AM_1^2r^2}{C(n-r)(n/2-r)}$ can be simplified to:

$$\sum_{p=1}^{y}(1-\eta C)^{(y-p)T_0}\frac{AM_1^2 r^2}{C(n-r)(n/2-r)} = \sum_{p=0}^{y-1}(1-\eta C)^{pT_0}\frac{AM_1^2 r^2}{C(n-r)(n/2-r)}$$

$$\leq \frac{1}{1-(1-\eta C)^{T_0}}\frac{AM_1^2 r^2}{C(n-r)(n/2-r)}.$$

So equation (37) can be further bounded as:

$$\begin{aligned}
&\|\mathbf{w}^I{}_{j_0+(y+m)T_0} - \mathbf{w}^U{}_{j_0+(y+m)T_0}\| \\
&\leq (1-\eta C)^{yT_0}\|\mathbf{w}^I{}_{j_0+mT_0} - \mathbf{w}^U{}_{j_0+mT_0}\| \\
&+ \frac{AM_1 r}{C(n-r)}(1-\eta C)^{yT_0}(1-\eta\mu)^{j_0}d_{0,mT_0-1}\frac{1}{1-(\frac{1-\eta\mu}{1-\eta C})^{T_0}} \\
&+ \frac{1}{1-(1-\eta C)^{T_0}}\frac{AM_1^2 r^2}{C(n-r)(n/2-r)}.
\end{aligned} \tag{38}$$

When $t \to \infty$ and thus $y \to \infty$, $(1-\eta C)^{yT_0} \to 0$ and thus

$$\|\mathbf{w}^I{}_{j_0+(y+m)T_0} - \mathbf{w}^U{}_{j_0+(y+m)T_0}\| = o(\frac{r}{n}).$$

□

A.3 Results for stochastic gradient descent

A.3.1 Quasi-Newton

We modify Equations (13) and (12) to SGD versions:

$$\mathbf{B}^{S}{}_{j_{k+1}} = \mathbf{B}^{S}{}_{j_k} - \frac{\mathbf{B}^{S}{}_{j_k}\Delta w^{S}{}_{j_k}\Delta w^{S}{}_{j_k}^{T}\mathbf{B}^{S}{}_{j_k}}{\Delta w^{S}{}_{j_k}^{T}\mathbf{B}^{S}{}_{j_k}\Delta w^{S}{}_{j_k}} + \frac{\Delta g^{S}{}_{j_k}\Delta g^{S}{}_{j_k}{}^{T}}{\Delta g^{S}{}_{j_k}{}^{T}\Delta w^{S}{}_{j_k}} \tag{39}$$

$$\mathbf{B}^{S^{-1}}{}_{j_{k+1}} = \left(\mathbf{I} - \frac{\Delta w^{S}{}_{j_k}\Delta g^{S}{}_{j_k}^{T}}{\Delta g^{S}{}_{j_k}^{T}\Delta w^{S}{}_{j_k}}\right)\mathbf{B}^{S^{-1}}{}_{j_k}\left(\mathbf{I} - \frac{\Delta g^{S}{}_{j_k}\Delta w^{S}{}_{j_k}^{T}}{\Delta g^{S}{}_{j_k}^{T}\Delta w^{S}{}_{j_k}}\right) + \frac{\Delta w^{S}{}_{j_k}\Delta w^{S}{}_{j_k}^{T}}{\Delta g^{S}{}_{j_k}^{T}\Delta w^{S}{}_{j_k}} \tag{40}$$

This iteration has the same initialization as $\mathbf{B}_{j_k}$ and $\mathbf{B}_{j_k}^{-1}$ but relies on the history information collected from the SGD-based training process $[\Delta w^{S}{}_{j_0}, \Delta w^{S}{}_{j_1}, \ldots, \Delta w^{S}{}_{j_{m-1}}]$ and $[\Delta g^{S}{}_{j_0}, \Delta g^{S}{}_{j_1}, \ldots, \Delta g^{S}{}_{j_{m-1}}]$ where $\Delta w^{S}{}_{j_x} = \mathbf{w}^{S}{}_{j_x} - \mathbf{w}^{I,S}{}_{j_x}$ and $\Delta g^{S}{}_{j_x} = G_{B,S}(\mathbf{w}^{I,S}{}_{j_x}) - G_{B,S}(\mathbf{w}^{S}{}_{j_x})$ $(x = 0, 1, 2, \ldots m-1)$. By the same argument as the proof of Lemma 6, the following inequality holds:

$$K_1\|\mathbf{z}\|^2 \leq \mathbf{z}^T\mathbf{B}^{S}{}_{j_k}\mathbf{z} \leq K_2\|\mathbf{z}\|^2 \tag{41}$$

where $K_1 := \frac{1}{(1+\frac{L}{\mu})^{2m}\frac{L}{\mu} + \frac{1-(1+\frac{L}{\mu})^{2m}}{1-(1+\frac{L}{\mu})^2}\frac{1}{\mu}}$ and $K_2 := (m+1)L$, which are both positive values representing a lower bound and an upper bound on the eigenvalues of $\mathbf{B}^{S}{}_{j_k}$.

A.3.2 Proof preliminaries

Similar to the argument for the GD-version of DeltaGrad, we can give an upper bound on $\delta_{t,S}$:

Lemma 9 (Upper bound on $\delta_{t,S}$)**.** *Define* $\delta_{t,S} = G_{B,S}(\boldsymbol{w}^{U,S}{}_t) - G^U_{B-\Delta B,S}(\boldsymbol{w}^{U,S}{}_t)$. *Then* $\|\delta_{t,S}\| \leq 2c_2\frac{\Delta B_t}{B}$. *Moreover, with probability higher than* $1 - t \times 2\exp(-2\sqrt{B})$,

$$\|\delta_{t',S}\| \leq 2c_2\left(\frac{r}{n} + \frac{1}{B^{1/4}}\right)$$

uniformly over all iterations $t' \leq t$.

Proof. Recall that

$$G_{B,S}(\mathbf{w}^{U,S}{}_t) = \frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla F_i(\mathbf{w}^{U,S}{}_t)$$

and

$$G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_t) = \frac{1}{B-\Delta B_t}\sum_{i\in\mathcal{B}_t, i\notin R}\nabla F_i(\mathbf{w}^{U,S}{}_t).$$

By subtracting $G_{B,S}(\mathbf{w}^{U,S}{}_t)$ from $G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_t)$, we have:

$$\begin{aligned}
&\|G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_t) - G_{B,S}(\mathbf{w}^{U,S}{}_t)\| \\
&= \|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla F_i(\mathbf{w}^{U,S}{}_t) - \frac{1}{B-\Delta B_t}\sum_{i\in\mathcal{B}_t, i\notin R}\nabla F_i(\mathbf{w}^{U,S}{}_t)\| \\
&= \|\frac{1}{B}\sum_{i\in\mathcal{B}_t, i\in R}\nabla F_i(\mathbf{w}^{U,S}{}_t) + (\frac{1}{B} - \frac{1}{B-\Delta B_t})\sum_{i\in\mathcal{B}_t, i\notin R}\nabla F_i(\mathbf{w}^{U,S}{}_t)\|
\end{aligned}$$

Then by using the triangle inequality and the fact that $\|\nabla F_i(\mathbf{w}^{U,S}{}_t)\| \leq c_2$ (Assumption 3), the formula above can be bounded by $\frac{2\Delta B_t c_2}{B}$.

Because of the randomness from SGD, the r removed samples can be viewed as uniformly distributed among all n training samples. Each sample is included in a mini-batch according to the outcome of a Bernoulli(r/n) random variable $\mathbf{S}_i$. Within a single mini-batch $\mathcal{B}_{t'}$ at the iteration t', we get $\mathbb{E}(\sum_{i\in\mathcal{B}_{t'}}\mathbf{S}_i) =$

$\mathbb{E}(\Delta B_{t'}) = B\frac{r}{n}$ and $\text{Var}(\sum_{i\in\mathcal{B}_{t'}} \mathbf{S}_i) = B\frac{r}{n}(1-\frac{r}{n})$. So in terms of the random variable $\frac{\Delta B_{t'}}{B}$, its expectation and variance will be $\mathbb{E}(\frac{\Delta B_{t'}}{B}) = \frac{r}{n}$ and $\text{Var}(\frac{\Delta B_{t'}}{B}) = \frac{r}{Bn}(1-\frac{r}{n})$.

Then based on Hoeffding's inequality, the following inequality holds:

$$\Pr(|\frac{\Delta B_{t'}}{B} - \frac{r}{n}|\leq \epsilon) \geq 1 - 2\exp(-2\epsilon^2 B).$$

Then by setting $\epsilon = \frac{1}{B^{1/4}}$ the formula above can be written as:

$$\Pr(|\frac{\Delta B_{t'}}{B} - \frac{r}{n}|\geq \frac{1}{B^{1/4}}) \leq 2\exp(-2\sqrt{B})$$

Then by taking the union for all the iterations before t, we get: with probability higher than $1 - t \times 2\exp(-2\sqrt{B})$,

$$|\frac{\Delta B_{t'}}{B} - \frac{r}{n}|\leq \frac{1}{B^{1/4}}$$

and thus

$$\frac{\Delta B_{t'}}{B} \leq \frac{r}{n} + \frac{1}{B^{1/4}} \tag{42}$$

for all $t' \leq t$. □

In what follows, we use Ψ_1 to represent $\Psi_1 := 2\exp(-2\sqrt{B})$, which goes to 0 with large B.

Next we provide a bound for the sum of random sampled Hessian matrices within a minibatch in SGD.

Theorem 6 (Hessian matrix bound in SGD)**.** *With probability higher than*

$$1 - 2p\exp\left(-\frac{\log(2p)\sqrt{B}}{4 + \frac{2}{3}\left(\frac{\log^2(2p)}{B}\right)^{1/4}}\right),$$

for a given iteration t, $\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_t} \boldsymbol{H}_i(\boldsymbol{w}^S{}_t)\right) - \boldsymbol{H}(\boldsymbol{w}^S{}_t)\| \leq L\left(\frac{\log^2(2p)}{B}\right)^{1/4}$ where p represents the number of model parameters.

Proof. We consider using the matrix Bernstein inequality, Lemma 5. We define the random matrix $\mathbf{S}_i = \frac{\mathbf{H}_i(\mathbf{w}) - \mathbf{H}(\mathbf{w})}{B}$ $(i \in \mathcal{B}_t)$. Due to the randomness from SGD, we know that $\mathbb{E}(\mathbf{S}_i) = \mathbb{E}(\frac{\mathbf{H}_i(\mathbf{w}) - \mathbf{H}(\mathbf{w})}{B}) = \mathbf{0}$. Using the sum $\mathbf{Z}$ as required in Lemma 5, $\mathbf{Z} = \left(\frac{1}{B}\sum_{i\in\mathcal{B}_t} \mathbf{H}_i(\mathbf{w})\right) - \mathbf{H}(\mathbf{w})$. Also note that $\mathbf{H}(\mathbf{w})$ and $\mathbf{H}_i(\mathbf{w})$ are both $p \times p$ matrices, so $d_1 = d_2 = p$ in Lemma 5.

Furthermore, for each $\mathbf{S}_i = \frac{\mathbf{H}_i(\mathbf{w}) - \mathbf{H}(\mathbf{w})}{B}$, its norm is bounded by $\frac{2L}{B}$ based on the smoothness condition, which means that $J = \frac{2L}{B}$ in Lemma 5. Then we can explicitly calculate the upper bound on $E(\mathbf{S}_i\mathbf{S}_i^*)$ and $V(\mathbf{Z})$:

$$\begin{aligned}
&\|\mathbb{E}(\mathbf{S}_i\mathbf{S}_i^*)\| \leq \mathbb{E}(\|\mathbf{S}_i\mathbf{S}_i^*\|) \leq \mathbb{E}(\|\mathbf{S}_i\|\|\mathbf{S}_i^*\|) \leq J^2 = \frac{4L^2}{B^2},\\
&V(Z) \leq \sum_{i\in\mathcal{B}_t} \frac{4L^2}{B^2} = \frac{4L^2}{B}.
\end{aligned}$$

Thus by plugging the above expression into equation (9) and (10), we get:

$$\begin{aligned}
P(\|\mathbf{Z}\| \geq x) &= \Pr(\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_t} \mathbf{H}_i(\mathbf{w})\right) - \mathbf{H}(\mathbf{w})\| \geq x)\\
&\leq (d_1 + d_2)\exp\left(\frac{-x^2}{\frac{4L^2}{B} + \frac{2Lx}{3B}}\right) = 2p\exp\left(\frac{-x^2}{\frac{4L^2}{B} + \frac{2Lx}{3B}}\right), \forall x \geq 0
\end{aligned} \tag{43}$$

$$\begin{aligned}\mathbb{E}(\|\mathbf{Z}\|) &= \mathbb{E}\left(\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_t}\mathbf{H}_i(\mathbf{w})\right)-\mathbf{H}(\mathbf{w})\|\right)\\ &\leq \sqrt{\frac{8L^2}{B}\log(d_1+d_2)}+\frac{2L}{3B}\log(d_1+d_2)=\sqrt{\frac{8L^2}{B}\log(2p)}+\frac{2L}{3B}\log(2p).\end{aligned} \tag{44}$$

Then by setting $x = L\left(\frac{\log^2(2p)}{B}\right)^{1/4}$, Equation (43) becomes:

$$\begin{aligned}&\Pr(\|\mathbf{Z}\|\geq L\left(\frac{\log^2(2p)}{B}\right)^{1/4})\\ &= Pr\left(\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_t}\mathbf{H}_i(\mathbf{w})\right)-\mathbf{H}(\mathbf{w})\|\geq L\left(\frac{\log^2(2p)}{B}\right)^{1/4}\right)\\ &\leq (2p)\exp\left(\frac{-\frac{L^2 log(2p)}{\sqrt{B}}}{\frac{4L^2}{B}+\frac{2L^2}{3B}\left(\frac{\log^2(2p)}{B}\right)^{1/4}}\right)=(2p)\exp\left(-\frac{\log(2p)\sqrt{B}}{4+\frac{2}{3}\left(\frac{\log^2(2p)}{B}\right)^{1/4}}\right).\end{aligned} \tag{45}$$

For large mini-batch size B, both $L\left(\frac{\log^2(2p)}{B}\right)^{1/4}$ and $(2p)\exp\left(-\frac{\log(2p)\sqrt{B}}{4+\frac{2}{3}\left(\frac{\log^2(2p)}{B}\right)^{1/4}}\right)$ are approaching 0. □

In what follows, we use Ψ_2 to denote the probability $\Psi_2 := (2p)\exp\left(-\frac{\log(2p)\sqrt{B}}{4+\frac{2}{3}\left(\frac{\log^2(2p)}{B}\right)^{1/4}}\right)$.

Based on this result, we can derive an SGD version of Theorem 1 as below, which also relies on a preliminary estimate on the bound on $\|\mathbf{w}^{I,S}{}_t - \mathbf{w}^S{}_t\|$:

Theorem 7 (Error in mean Hessian, and in secant equation with incorrect quasi-Hessian for SGD)**.** *Suppose that* $\|\boldsymbol{w}^S{}_{t'} - \boldsymbol{w}^{I,S}{}_{t'}\|\leq M_1\frac{1}{\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}}}(\frac{r}{n}+\frac{1}{B^{1/4}})$ *and*

$$\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_{t'}}\boldsymbol{H}_i(\boldsymbol{w}^S{}_{t'})\right)-\boldsymbol{H}(\boldsymbol{w}^S{}_{t'})\|\leq L\left(\frac{\log^2(2p)}{B}\right)^{1/4}$$

hold for any $t'\leq t$ *where* $M_1=\frac{2c_2}{\mu}$, μ *is from Assumption 2 and* c_2 *is from Assumption 3. Let* $e=\frac{L(L+1)+K_2L}{\mu K_1}$ *for the upper and lower bounds* K_1, K_2 *on the eigenvalues of the quasi-Hessian from equation* (41) *and for the Lipshitz constant* c_0 *of the Hessian. For any* t_1, t_2 *such that* $1\leq t_1<t_2\leq t$*, we have:*

$$\|\boldsymbol{H}^S{}_{t_1}-\boldsymbol{H}^S{}_{t_2}\|\leq 2L\left(\frac{\log^2(2p)}{B}\right)^{1/4}+c_0d_{t_1,t_2}+3c_0M_1\frac{1}{\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}}}(\frac{r}{n}+\frac{1}{B^{1/4}}).$$

For any $j_1, j_2, \ldots, j_m$ *such that* $j_m\leq t'\leq j_m+T_0-1$ *and* $t'\leq t$*, we have:*

$$\|\Delta g^S{}_{j_k}-\boldsymbol{B}^S{}_{j_q}\Delta w^S{}_{j_k}\|$$

$$\leq\left[(1+e)^{j_q-j_k-1}-1\right]\cdot[2L\left(\frac{\log^2(2p)}{B}\right)^{1/4}+c_0d_{j_k,j_q}+\frac{3c_0M_1}{\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}}}(\frac{r}{n}+\frac{1}{B^{1/4}})]\cdot s_{j_m,j_1}.$$

Here $s_{j_m,j_1}=\max\left(\|\Delta w^S{}_a\|\right)_{a=j_1,j_2,\ldots,j_m}$, $d_{j_k,j_q}=\max\left(\|\boldsymbol{w}^S{}_a-\boldsymbol{w}^S{}_b\|\right)_{j_k\leq a\leq b\leq j_q}$, $\boldsymbol{H}^S{}_t$ *is the average of the Hessian matrix evaluated between* $\boldsymbol{w}^S{}_t$ *and* $\boldsymbol{w}^{I,S}{}_t$ *for the samples in mini-batch* $\mathcal{B}_t$*:*

$$\boldsymbol{H}^S{}_t=\frac{1}{B}\sum_{i\in\mathcal{B}_t}\int_0^1\boldsymbol{H}_i(\boldsymbol{w}^S{}_t+x(\boldsymbol{w}^{U,S}{}_t-\boldsymbol{w}^S{}_t))dx.$$

Proof. First of all, let us bound $\|\mathbf{H}^{S}{}_{t_1} - \int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}))dx\|$ by adding and subtracting $\frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1})$ and $\mathbf{H}(\mathbf{w}^{S}{}_{t_1})$ inside the norm:

$$
\begin{aligned}
&\|\mathbf{H}^{S}{}_{t_1} - \int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}))dx\| \\
&= \|\int_0^1 \frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}))dx - \int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}))dx\| \\
&= \|\int_0^1 \frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} (\mathbf{H}_i(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1})) - \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1}))dx + \frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1}) \\
&- \int_0^1 (\mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1})) - \mathbf{H}(\mathbf{w}^{S}{}_{t_1}))dx - \mathbf{H}(\mathbf{w}^{S}{}_{t_1})\|.
\end{aligned}
$$

Then by using the triangle inequality and Assumption 4, the formula above can be bounded as:

$$
\begin{aligned}
&\leq \int_0^1 \frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \|\mathbf{H}_i(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1})) - \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1})\|dx \\
&+ \int_0^1 \|\mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1})) - \mathbf{H}(\mathbf{w}^{S}{}_{t_1})\|dx \\
&+ \|\int_0^1 \frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1}) - \mathbf{H}(\mathbf{w}^{S}{}_{t_1})\| \\
&\leq \frac{1}{B}(\sum_{i\in\mathcal{B}_{t_1}} \int_0^1 c_0 x\|\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}\|dx) + \int_0^1 c_0 x\|\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}\|dx \\
&+ \|\frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1}) - \mathbf{H}(\mathbf{w}^{S}{}_{t_1})\| \\
&\leq c_0\|\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}\| + \|\frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1}) - \mathbf{H}(\mathbf{w}^{S}{}_{t_1})\|.
\end{aligned}
\tag{46}
$$

Then based on the above results, we can compute the bound on $\|\mathbf{H}^{S}{}_{t_1} - \mathbf{H}^{S}{}_{t_2}\|$, for which we use the triangle inequality first:

$$
\begin{aligned}
&\|\mathbf{H}^{S}{}_{t_1} - \mathbf{H}^{S}{}_{t_2}\| \\
&= \|\mathbf{H}^{S}{}_{t_1} - \int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}))dx\| \\
&+ \|\int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}))dx - \int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_2} + x(\mathbf{w}^{I,S}{}_{t_2} - \mathbf{w}^{S}{}_{t_2}))dx\| \\
&+ \|\int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_2} + x(\mathbf{w}^{I,S}{}_{t_2} - \mathbf{w}^{S}{}_{t_2}))dx - \mathbf{H}^{S}{}_{t_2}\|.
\end{aligned}
\tag{47}
$$

Then by using the result from Formula (46), this term can be further bounded as:

$$
\begin{aligned}
&\leq c_0\|\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}\| + \|\int_0^1 \frac{1}{B}\sum_{i\in\mathcal{B}_{t_1}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_1}) - \mathbf{H}(\mathbf{w}^{S}{}_{t_1})\| \\
&+ c_0\|\mathbf{w}^{I,S}{}_{t_2} - \mathbf{w}^{S}{}_{t_2}\| + \|\int_0^1 \frac{1}{B}\sum_{i\in\mathcal{B}_{t_2}} \mathbf{H}_i(\mathbf{w}^{S}{}_{t_2}) - \mathbf{H}(\mathbf{w}^{S}{}_{t_2})\| \\
&+ \|\int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_1} + x(\mathbf{w}^{I,S}{}_{t_1} - \mathbf{w}^{S}{}_{t_1}))dx - \int_0^1 \mathbf{H}(\mathbf{w}^{S}{}_{t_2} + x(\mathbf{w}^{I,S}{}_{t_2} - \mathbf{w}^{S}{}_{t_2}))dx\|.
\end{aligned}
$$

Since

$$\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_{t'}}\mathbf{H}_i(\mathbf{w}^S{}_{t'})\right)-\mathbf{H}(\mathbf{w}^S{}_{t'})\|\leq L\left(\frac{\log^2(2p)}{B}\right)^{1/4}$$

for any $t'\leq t$, then the formula above can be bounded as:

$$\begin{aligned}
&\leq 2L\left(\frac{\log^2(2p)}{B}\right)^{1/4}+c_0\|\mathbf{w}^S{}_{t_1}-\mathbf{w}^S{}_{t_2}\|+\frac{c_0}{2}\|\mathbf{w}^S{}_{t_1}-\mathbf{w}^{I,S}{}_{t_1}\|\\
&+\frac{c_0}{2}\|\mathbf{w}^{I,S}{}_{t_2}-\mathbf{w}^S{}_{t_2}\|+c_0\|\mathbf{w}^{I,S}{}_{t_1}-\mathbf{w}^S{}_{t_1}\|+c_0\|\mathbf{w}^{I,S}{}_{t_2}-\mathbf{w}^S{}_{t_2}\|\\
&=2L\left(\frac{\log^2(2p)}{B}\right)^{1/4}+3c_0M_1\frac{1}{\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}}}(\frac{r}{n}+\frac{1}{B^{1/4}})+c_0d_{t_1,t_2}.
\end{aligned}$$

This finishes the proof of the first inequality. Then by defining

$$f=\left(2L\left(\frac{\log^2(2p)}{B}\right)^{1/4}+3c_0M_1\frac{1}{\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}}}(\frac{r}{n}+\frac{1}{B^{1/4}})+c_0d_{j_k,j_q}\right)s_{j_m,j_1}$$

and using the same argument as Equation (21)-(23) (except that Δw and Δg are replaced with Δw^S and Δg^S), the following inequality thus holds:

$$\begin{aligned}
b_{j_q}&=\|\Delta g^S{}_{j_k}-\left(\mathbf{B}^S{}_{j_q}-\frac{\mathbf{B}^S{}_{j_q}\Delta w^S{}_{j_q}\Delta w^{S^T}_{j_q}\mathbf{B}^S{}_{j_q}}{\Delta w^{S^T}_{j_q}\mathbf{B}^S{}_{j_q}\Delta w^S{}_{j_q}}+\frac{\Delta g^S{}_{j_q}\Delta g^{S^T}_{j_q}}{\Delta g^{S^T}_{j_q}\Delta w^S{}_{j_q}}\right)\Delta w^S{}_{j_k}\| \\
&\leq[(1+e)^{j_q-j_k}-1]f
\end{aligned}\tag{48}$$

and thus

$$\|\Delta g^S{}_{j_k}-\mathbf{B}^S{}_{j_q}\Delta w^S{}_{j_k}\|\leq[(1+e)^{j_q-j_k-1}-1]f,$$

which finishes the proof.

For simplicity, we denote $M_1^S:=M_1\frac{1}{\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}}}$. So the preliminary estimate of the bound on $\|\mathbf{w}^S{}_{t'}-\mathbf{w}^{I,S}{}_{t'}\|$ becomes: $\|\mathbf{w}^S{}_{t'}-\mathbf{w}^{I,S}{}_{t'}\|\leq M_1^S(\frac{r}{n}+\frac{1}{B^{1/4}})$

□

Similarly, we get a SGD-version of Corollary 1:

Corollary 2 (Approximation accuracy of Quasi-Hessian to mean Hessian)**.** *Suppose that* $\|\boldsymbol{w}^S{}_{t'}-\boldsymbol{w}^{I,S}{}_{t'}\|\leq M_1^S(\frac{r}{n}+\frac{1}{B^{1/4}})$ *and*

$$\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_{t'}}\boldsymbol{H}_i(\boldsymbol{w}^S{}_{t'})\right)-\boldsymbol{H}(\boldsymbol{w}^S{}_{t'})\|\leq L\left(\frac{\log^2(2p)}{B}\right)^{1/4}$$

hold for any $t'\leq t$. M_1 *and* M_1^S *are provided in Theorem 7, i.e.* $M_1=\frac{2c_2}{\mu}$ *and* $M_1^S=M_1\frac{1}{\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}}}$. *Then for any* t' *and* j_m *such that* $j_m\leq t'\leq j_m+T_0-1$ *and* $t'\leq t$, *the following inequality holds:*

$$\begin{aligned}
&\|\boldsymbol{H}^S{}_{t'}-\boldsymbol{B}^S{}_{j_m}\|\leq\xi^S_{j_1,j_m}\\
&:=A(d_{j_1,j_m+T_0-1}+3M_1^S(\frac{r}{n}+\frac{1}{B^{1/4}})+\frac{2}{c_0}L\left(\frac{\log^2(2p)}{B}\right)^{1/4})
\end{aligned}$$

where recall again that c_0 *is the Lipschitz constant of the Hessian,* d_{j_1,j_m+T_0-1} *is the maximal gap between the iterates of the SGD algorithm on the full data from* j_1 *to* j_m+T_0-1 *and* $A=\frac{c_0\sqrt{m}[(1+e)^m-1]}{c_1}+c_0$ *in which* e *is a problem dependent constant defined in Theorem 7,* c_1 *is the "strong independence" constant from Assumption* (5).

This proof is similar to the proof of Corollary 1. First of all, $\mathbf{H}$, $\mathbf{B}$, ξ_{j_1,j_m} in Corollary 1 are replaced with ${\mathbf{H}^S}$, ${\mathbf{B}^S}$, $\xi^S_{j_1,j_m}$. Second, Theorem 7 holds and thus the following inequality holds:

$$\|{\mathbf{H}^S}_{t'} - {\mathbf{H}^S}_{j_m}\| \leq 2L\left(\frac{\log^2(2p)}{B}\right)^{1/4} + c_0 d_{t',j_m} + 3c_0 M_1^S(\frac{r}{n} + \frac{1}{B^{1/4}})$$

by using strong independence from Assumption 5, $\|{\mathbf{H}^S}_{j_m} - {\mathbf{B}^S}_{j_m}\|$ can be bounded as:

$$\|{\mathbf{H}^S}_{j_m} - {\mathbf{B}^S}_{j_m}\| \leq \sqrt{m}[(1+e)^m - 1]\frac{c_0}{c_1} \cdot \left(d_{j_1,j_m+T_0-1} + 3M_1^S(\frac{r}{n} + \frac{1}{B^{1/4}}) + \frac{2}{c_0}L\left(\frac{\log^2(2p)}{B}\right)^{1/4}\right) \quad (49)$$

Then by combining the two formulas above, we know that Corollary 2 holds. Note that the definition of $\xi^S_{j_1,j_m}$ can be rewritten as below:

$$\begin{aligned}\xi^S_{j_1,j_m} &= A(d_{j_1,j_m+T_0-1} + 3M_1^S(\frac{r}{n} + \frac{1}{B^{1/4}}) + \frac{2}{c_0}L\left(\frac{\log^2(2p)}{B}\right)^{1/4}) \\ &=: Ad_{j_1,j_m+T_0-1} + A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}}\end{aligned} \quad (50)$$

in which $A_1 := 3AM_1^S$ and $A_2 := 3AM_1^S + \frac{2AL(log(2p))^{1/2}}{c_0}$.

We can do a similar analysis to Lemma 8 by simply replacing $\mathbf{w}_t$ and $F(*)$ with ${\mathbf{w}^S}_t$ and $G_{B,S}$:

Lemma 10. *Let us use the definition of $d_{k,q}$ from Theorem 7:*

$$d_{k,q} = \max\left(\|{\boldsymbol{w}^S}_a - {\boldsymbol{w}^S}_b\|\right)_{k\leq a\leq b\leq q}$$

where $k < q \leq t$, then $d_{k,q} \leq (1-\eta\mu)^k d_{0,q-j} + 2c_2\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}$ holds with probability higher than $1 - t(p+1)\exp\left(-\frac{\log(p+1)\sqrt{B}}{4+\frac{2}{3}\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}\right)$.

Proof. According to Lemma 5, we can define a random matrix $\mathbf{S}_i = \frac{1}{B}(\nabla F_i({\mathbf{w}^S}_a) - \nabla F({\mathbf{w}^S}_a)$ where recall that $\nabla F({\mathbf{w}^S}_a) = \frac{1}{n}\sum_{i=1}^n \nabla F_{i,a}({\mathbf{w}^S}_a)$ $(i \in \mathcal{B}_t)$. Due to the randomness from SGD, we know that $\mathbb{E}(\mathbf{S}_i) = \mathbf{0}$. Based on the definition of $\mathbf{Z}$ in Lemma 5, $\mathbf{Z} = \frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla F_i({\mathbf{w}^S}_a) - \nabla F({\mathbf{w}^S}_a)$. Also note that $\nabla F_i({\mathbf{w}^S}_a)$ and $\nabla F({\mathbf{w}^S}_a)$ are both $p \times 1$ matrices, so $d_1 = p$ and $d_2 = 1$ in Lemma 5.

Moreover according to Assumption 3, $\|\nabla F_i({\mathbf{w}^S}_a)\| \leq c_2$. Then we know that $V(\mathbf{Z}) \leq \frac{4c_2^2}{B}$ and $\|\mathbf{S}_i\| \leq \frac{2c_2}{B}$. So according to Lemma 5, the following inequality holds:

$$\begin{aligned}P(\|\mathbf{Z}\| \geq x) &= \Pr(\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla F_i({\mathbf{w}^S}_a) - \nabla F({\mathbf{w}^S}_a)\| \geq x) \\ &\leq (d_1+d_2)\exp\left(\frac{-x^2}{\frac{4c_2^2}{B} + \frac{2c_2 x}{3B}}\right) = (p+1)\exp\left(\frac{-x^2}{\frac{4c_2^2}{B} + \frac{2c_2 x}{3B}}\right), \forall x \geq 0\end{aligned} \quad (51)$$

By setting $x = c_2\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}$, the formula above is evaluated as:

$$\begin{aligned}&\Pr(\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla F_i({\mathbf{w}^S}_a) - \nabla F({\mathbf{w}^S}_a)\| \geq c_2\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}) \\ &\leq (p+1)\exp\left(-\frac{\log(p+1)\sqrt{B}}{4+\frac{2}{3}\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}\right)\end{aligned}$$

So by taking the union for the first t iterations, then with probability higher than $1-t(p+1)\exp\left(-\frac{\log(p+1)\sqrt{B}}{4+\frac{2}{3}\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}\right)$, the following inequality holds for all $t' \leq t$:

$$\|\frac{1}{B}\sum_{i\in\mathcal{B}_{t'}}\nabla F_i(\mathbf{w}^S{}_a) - \nabla F(\mathbf{w}^S{}_a)\| \leq c_2\left(\frac{(\log(p+1))^2}{B}\right)^{1/4} \tag{52}$$

Then by using the similar arguments to Lemma 8, we get:

$$\|\mathbf{w}^S{}_a - \mathbf{w}^S{}_b\| \leq (1-\eta\mu)^z\|\mathbf{w}^S{}_{a-z} - \mathbf{w}^S{}_{b-z}\| + \frac{2c_2}{\mu}\left(\frac{(\log(p+1))^2}{B}\right)^{1/4} = (1-\eta\mu)^z\|\mathbf{w}^S{}_{a-z} - \mathbf{w}^S{}_{b-z}\| + M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}$$

and thus $d_{k,q} \leq (1-\eta\mu)^k d_{0,q-k} + M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}$ holds with probability higher than $1-t(p+1)\exp\left(-\frac{\log(p+1)\sqrt{B}}{4+\frac{2}{3}\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}\right)$.

In what follows, we use Ψ_3 to denote $(p+1)\exp\left(-\frac{\log(p+1)\sqrt{B}}{4+\frac{2}{3}\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}\right)$.

□

Then by using the definition of $\xi^S_{j_1,j_m}$, the following inequality holds with probability higher than $1-t\Psi_3$ for any x such that for $j_0 + (x+m-1)T_0 \leq t$, the following inequality holds:

$$\begin{aligned}\xi^S_{j_1,j_m} &= \xi^S_{j_0+xT_0,j_0+(x+m-1)T_0} \leq (1-\eta\mu)^{xT_0}Ad_{j_0,j_0+mT_0-1} \\ &+ A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}\end{aligned} \tag{53}$$

A.3.3 Main recursions

We bound the difference between $\mathbf{w}^{I,S}{}_t$ and $\mathbf{w}^{U,S}{}_t$. First we bound $\|\mathbf{w}^S{}_t - \mathbf{w}^{U,S}{}_t\|$:

Theorem 8 (Bound between iterates on full and the leave-r-out dataset)**.** *When*

$$\frac{\Delta B_{t'}}{B} \leq \frac{r}{n} + \frac{1}{B^{1/4}}$$

holds for all $t' < t$, $\|\boldsymbol{w}^S{}_t - \boldsymbol{w}^{U,S}{}_t\| \leq \frac{2c_2}{\mu}(\frac{r}{n} + \frac{1}{B^{1/4}})$. Since with probability higher than $1 - t \times \Psi_1$,

$$\frac{\Delta B_{t'}}{B} \leq \frac{r}{n} + \frac{1}{B^{1/4}}$$

holds for all $t' < t$. Then with the same probability, $\|\boldsymbol{w}^S{}_{t'+1} - \boldsymbol{w}^{U,S}{}_{t'+1}\| \leq M_1(\frac{r}{n} + \frac{1}{B^{1/4}})$ for all iterations $t' < t$, where recall that $M_1 = \frac{2c_2}{\mu}$.

Similarly, we can bound the difference between $\mathbf{w}^I{}_t$ and $\mathbf{w}_t$.

Theorem 9 (Bound between iterates on full data and incrementally updated ones)**.** *Suppose that for at some iteration t and any given $t' \leq t$ such that $j'_m \leq t' \leq j'_m + T_0 - 1$, we have the following bounds:*

1. $\|\boldsymbol{H}^S{}_{t'} - \boldsymbol{B}^S{}_{j'_m}\| \leq \xi^S_{j'_1,j'_m} = Ad_{j'_1,j'_m+T_0-1} + A\frac{3}{\frac{1}{2}-\frac{r}{n}}M_1(\frac{r}{n} + \frac{1}{B^{1/4}}) + A\frac{2}{c_0}L\left(\frac{\log^2(2p)}{B}\right)^{1/4}$;
2. $\frac{\Delta B_{t'}}{B} \leq \frac{r}{n} + \frac{1}{B^{1/4}}$;
3. *Formula* (53) *holds for any x such that $j_0 + (x+m-1)T_0 \leq t$;*
4. $\xi^S_{j_0,j_0+(m-1)T_0} + A \times M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4} \leq \frac{\mu}{2}$,

then

$$\|\boldsymbol{w}^{I,S}{}_{t'+1} - \boldsymbol{w}^{S}{}_{t'+1}\| \leq \frac{2c_2}{(\frac{1}{2} - \frac{r}{n} - \frac{1}{B^{1/4}})\mu}(\frac{r}{n} + \frac{1}{B^{1/4}}) = M_1^S(\frac{r}{n} + \frac{1}{B^{1/4}})$$

for any $t' \leq t$ Recall that c_0 is the Lipshitz constant of the Hessian, M_1 and A are defined in Theorem 8 and Corollary 2 respectively, and do not depend on t.

in particular for all t, the following inequality holds:

$$\|\boldsymbol{w}^{I,S}{}_{t+1} - \boldsymbol{w}^{S}{}_{t+1}\| \leq \frac{2c_2}{(\frac{1}{2} - \frac{r}{n} - \frac{1}{B^{1/4}})\mu}(\frac{r}{n} + \frac{1}{B^{1/4}}) = M_1^S(\frac{r}{n} + \frac{1}{B^{1/4}}).$$

Similarly, we will show that both inequalities $\|\mathbf{H}^{S}{}_{t} - \mathbf{B}^{S}{}_{j_m}\| \leq \xi^S_{j_1,j_m}$ and $\|\mathbf{w}^{I,S}{}_{t+1} - \mathbf{w}^{S}{}_{t+1}\| \leq M_1^S(\frac{r}{n} + \frac{1}{B^{1/4}})$ hold for all iterations t.

Theorem 10 (Bound between iterates on full data and incrementally updated ones (all iterations))**.** *Suppose that there are T iterations in total for each training phase, then with probability higher than $1 - T \times (\Psi_1 + \Psi_2 + \Psi_3)$, for any t where $j_m < t < j_m + T_0 - 1$, $\|\boldsymbol{w}^{I,S}{}_{t} - \boldsymbol{w}^{S}{}_{t}\| \leq \frac{1}{\frac{1}{2} - \frac{r}{n} - \frac{1}{B^{1/4}}} M_1(\frac{r}{n} + \frac{1}{B^{1/4}})$ and $\|\boldsymbol{H}^{S}{}_{t} - \boldsymbol{B}^{S}{}_{j_m}\| \leq \xi^S_{j_1,j_m}$, where $\xi^S_{j_1,j_m}$ is defined in Corollary 2, Ψ_1 is defined in Lemma 9, Ψ_2 is defined in Theorem 6 and Ψ_3 is defined in Lemma 10.*

Then we have the following bound for $\|\mathbf{w}^{U}{}_{t} - \mathbf{w}^{I}{}_{t}\|$.

Theorem 11 (Main result: Bound between true and incrementally updated iterates for SGD)**.** *Suppose that there are T iterations in total for each training phase, then with probability higher than $1 - T \times (\Psi_1 + \Psi_2 + \Psi_3)$, the result $\boldsymbol{w}^{I,S}{}_{t}$ of Algorithm 1 approximates the correct iteration values $\boldsymbol{w}^{U,S}{}_{t}$ at the rate*

$$\|\boldsymbol{w}^{U,S}{}_{t} - \boldsymbol{w}^{I,S}{}_{t}\| \leq o((\frac{r}{n} + \frac{1}{B^{1/4}})).$$

So $\|\boldsymbol{w}^{U,S}{}_{t} - \boldsymbol{w}^{I,S}{}_{t}\|$ is of a lower order than $(\frac{r}{n} + \frac{1}{B^{1/4}})$.

A.3.4 Proof of Theorem 8

Proof. By subtracting $\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t}$, taking the matrix norm and using the update rule in equation (5) and (6), we get:

$$\begin{aligned}
&\|\mathbf{w}^{S}{}_{t+1} - \mathbf{w}^{U,S}{}_{t+1}\| \\
&= \|\mathbf{w}^{S}{}_{t} - \eta G_{B,S}(\mathbf{w}^{S}{}_{t}) - \left(\mathbf{w}^{U,S}{}_{t} - \eta G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_{t})\right)\| \\
&= \|\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t} - \eta\left(G_{B,S}(\mathbf{w}^{S}{}_{t}) - G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_{t})\right)\| \\
&= \|\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t} - \eta(G_{B,S}(\mathbf{w}^{S}{}_{t}) - G_{B,S}(\mathbf{w}^{U,S}{}_{t}) \\
&+ G_{B,S}(\mathbf{w}^{U,S}{}_{t}) - G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_{t}))\| \\
&= \|\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t} - \eta\left(G_{B,S}(\mathbf{w}^{S}{}_{t}) - G_{B,S}(\mathbf{w}^{U,S}{}_{t})\right) + \\
&\eta\left(G_{B,S}(\mathbf{w}^{U,S}{}_{t}) - G^U_{B-\Delta B,S}(\mathbf{w}^{U,S}{}_{t})\right)\|
\end{aligned} \tag{54}$$

By Cauchy mean-value theorem and the triangle inequality, the above formula becomes:

$$\leq \|\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t} - \eta(\frac{1}{B}\int_0^1 \sum_{i\in\mathcal{B}_t} \mathbf{H}_i\left(\mathbf{w}^{S}{}_{t} + x\left(\mathbf{w}^{U,S}{}_{t} - \mathbf{w}^{S}{}_{t}\right)\right)dx)\left(\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t}\right)\| + \eta\|\delta_{t,S}\|$$

$$= \|\left(\mathbf{I} - \eta(\frac{1}{B}\int_0^1 \sum_{i\in\mathcal{B}_t} \mathbf{H}_i\left(\mathbf{w}^{S}{}_{t} + x\left(\mathbf{w}^{U,S}{}_{t} - \mathbf{w}^{S}{}_{t}\right)\right)dx)\right)\left(\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t}\right)\| + \eta\|\delta_{t,S}\|$$

Then by using the Lemma 3 and Lemma 9, the formula above can be bounded as:

$$\leq (1 - \eta\mu)\|\mathbf{w}^{S}{}_{t} - \mathbf{w}^{U,S}{}_{t}\| + \eta 2c_2\frac{\Delta B_t}{B}$$

Then by using Lemma 9 and using the formula above recursively, we get that with probability higher than $1 - t \cdot \Psi_1$, $\|\mathbf{w}^S{}_{t'} - \mathbf{w}^{U,S}{}_{t'}\| \leq \frac{2c_2}{\mu}(\frac{r}{n} + \frac{1}{B^{1/4}})$ holds for all iterations $t' \leq t$, which finishes the proof. □

A.3.5 Proof of Theorem 9

Proof. For any $t' \leq t$, by subtracting $\mathbf{w}^S{}_{t'}$ by $\mathbf{w}^{I,S}{}_{t'}$ and taking the same argument as equation (29)-(31) (except that $\mathbf{w}_{t'}$, $\mathbf{w}^I{}_{t'}$, $\mathbf{H}$, $\mathbf{B}$, n, r are replaced with $\mathbf{w}^S{}_{t'}$, $\mathbf{w}^{I,S}{}_{t'}$, $\mathbf{H}^S$, $\mathbf{B}^S$, B, $\Delta B_{t'}$), the following equality holds due to the bound on $\|\mathbf{H}^S{}_{t'} - \mathbf{B}^S{}_{j_m}\|$:

$$\begin{aligned}&\|\mathbf{w}^{I,S}{}_{t'+1} - \mathbf{w}^S{}_{t'+1}\| \\ &\leq (1 - \eta\mu + \eta\frac{B}{B - \Delta B_t}\xi^S_{j_1,j_m})\|\mathbf{w}^{I,S}{}_t - \mathbf{w}^S{}_t\| + \frac{2\Delta B_t \eta c_2}{B - \Delta B_t}.\end{aligned} \tag{55}$$

Since $\frac{\Delta B_t}{B} \leq \frac{r}{n} + \frac{1}{B^{1/4}}$ for all iterations between 0 and t, the following two inequalities hold:

$$\frac{2\Delta B_t \eta c_2}{B - \Delta B_t} = \frac{2\eta c_2}{\frac{B}{\Delta B_t} - 1} \leq \frac{2\eta c_2}{\frac{1}{\frac{r}{n} + \frac{1}{B^{1/4}}} - 1} = \frac{2\eta c_2}{1 - \frac{r}{n} - \frac{1}{B^{1/4}}}(\frac{r}{n} + \frac{1}{B^{1/4}}), \tag{56}$$

$$\frac{B}{B - \Delta B_t} = \frac{1}{1 - \frac{\Delta B_t}{B}} \leq \frac{1}{1 - (\frac{r}{n} + \frac{1}{B^{1/4}})}. \tag{57}$$

Moreover, since Formula (53) holds and $\xi^S_{j_0,j_0+(m-1)T_0} + A \times M_1 \left(\frac{(\log(p+1))^2}{B}\right)^{1/4} \leq \frac{\mu}{2}$, then:

$$\begin{aligned}&\xi^S_{j_1,j_m} = \xi^S_{j_0+xT_0,j_0+(x+m-1)T_0} \leq (1 - \eta\mu)^{xT_0} A d_{j_0,j_0+mT_0-1} \\ &+ A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4} \\ &\leq A d_{j_0,j_0+mT_0-1} + A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4} \\ &= \xi_{j_0,j_0+mT_0-1} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4} \leq \frac{\mu}{2}.\end{aligned}$$

Then the Formula (55) can be bounded as:

$$\begin{aligned}&\|\mathbf{w}^{I,S}{}_{t'+1} - \mathbf{w}^S{}_{t'+1}\| \\ &\leq (1 - \eta\mu + \eta\frac{B}{B - \Delta B_t}(\xi^S_{j_0,j_0+(m-1)T_0} + A \times M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4})\|\mathbf{w}^{I,S}{}_{t'} - \mathbf{w}^S{}_{t'}\| + \frac{2\Delta B_t \eta c_2}{B - \Delta B_t} \\ &\leq (1 - \eta\mu + \eta\frac{\xi^S_{j_0,j_0+(m-1)T_0} + A \times M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}{1 - (\frac{r}{n} + \frac{1}{B^{1/4}})})\|\mathbf{w}^{I,S}{}_{t'} - \mathbf{w}^S{}_{t'}\| + \frac{2\eta c_2}{1 - \frac{r}{n} - \frac{1}{B^{1/4}}}(\frac{r}{n} + \frac{1}{B^{1/4}}),\end{aligned}$$

which uses equation (56) and (57). Then applying the formula recursively from iteration t to 0, we can get:

$$\begin{aligned}&\|\mathbf{w}^{I,S}{}_{t'+1} - \mathbf{w}^S{}_{t'+1}\| \\ &\leq \frac{1}{\eta(\mu - \frac{\xi^S_{j_0,j_0+(m-1)T_0} + 2c_2\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}{1 - (\frac{r}{n} + \frac{1}{B^{1/4}})})} \frac{2\eta c_2}{1 - \frac{r}{n} - \frac{1}{B^{1/4}}}(\frac{r}{n} + \frac{1}{B^{1/4}}).\end{aligned}$$

Then since $\xi^S_{j_0,j_0+(m-1)T_0} \leq \frac{\mu}{2}$, the formula above can be further bounded as:

$$= \frac{2c_2}{(1-\frac{r}{n}-\frac{1}{B^{1/4}})\mu - \xi^S_{j_0,j_0+(m-1)T_0} - A \times M_1 \left(\frac{(\log(p+1))^2}{B}\right)^{1/4}}(\frac{r}{n}+\frac{1}{B^{1/4}})$$
$$\leq \frac{2c_2}{(\frac{1}{2}-\frac{r}{n}-\frac{1}{B^{1/4}})\mu}(\frac{r}{n}+\frac{1}{B^{1/4}}).$$

□

A.3.6 Proof of Theorem 10

The proof is the same as the proof of Theorem 4 except that $\mathbf{w}$, $\mathbf{w}^I$, n, r, ξ_{j_1,j_m}, $\mathbf{H}$, $\mathbf{B}$ need to be replaced by $\mathbf{w}^S$, $\mathbf{w}^{I,S}$, B, r, $\xi^S_{j_1,j_m}$, $\mathbf{H}^S$, $\mathbf{B}^S$ and the main theorems that the proof depends on will be replaced by Theorem 9 and Corollary 2. But we need some careful explanations for the probability, which is shown as:

Proof. We define the following event at a given iteration k:

$$\Omega_0(k) = \{\|{\mathbf{w}^S}_k - {\mathbf{w}^{U,S}}_k\| \leq M_1(\frac{r}{n}+\frac{1}{B^{1/4}})\},$$
$$\Omega_1(k) = \{\|{\mathbf{w}^S}_k - {\mathbf{w}^{I,S}}_k\| \leq M_1^S(\frac{r}{n}+\frac{1}{B^{1/4}})\},$$
$$\Omega_2(k) = \{\|{\mathbf{H}^S}_{k-1} - {\mathbf{B}^S}_{j_m}\| \leq \xi^S_{j_1,j_m}\} \quad (j_m \leq k-1 \leq j_m+T_0-1),$$
$$\Omega_3(k) = \{\|\left(\frac{1}{B}\sum_{i\in\mathcal{B}_{k-1}} \mathbf{H}_i({\mathbf{w}^S}_{k-1})\right) - \mathbf{H}({\mathbf{w}^S}_{k-1})\| \leq L\left(\frac{\log^2(2p)}{B}\right)^{1/4}\},$$
$$\Omega_4(k) = \{\xi^S_{j_0+xT_0,j_0+(x+m-1)T_0} \leq (1-\eta\mu)^{j_0+xT_0}Ad_{0,mT_0-1}$$
$$+ A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}\} \text{ where } j_0+(x+m-1)T_0 \leq k-1 \leq j_0+(x+m)T_0-1,$$
$$\Omega_5(k) = \{\frac{\Delta B_{k-1}}{B} \leq \frac{r}{n}+\frac{1}{B^{1/4}}\}.$$

For all t, according to Corollary 2, the following equation holds:

$$\Pr(\bigcap_{k=1}^{t}\Omega_2(k) | \bigcap_{k=1}^{t-1}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_3(k)) = 1.$$

in which the co-occurrence of multiple events is denoted by $\bigcap$ or ",". So this formula means that the probability that $\Omega_2(k)$ is true for all $k \leq t$ given that the events $\Omega_1(k)$ and $\Omega_3(k)$ are true at the same time for all $k \leq t$ is 1.

Similarly, according to Theorem 9, $\Pr(\bigcap_{k=1}^{t}\Omega_1(k) \Big| \bigcap_{k=1}^{t}\Omega_2(k), \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k)) = 1$. Then we know that:

$$\Pr(\bigcap_{k=1}^{t}\Omega_1(k) \Big| \bigcap_{k=1}^{t}\Omega_2(k), \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k)) \cdot \Pr(\bigcap_{k=1}^{t}\Omega_2(k) | \bigcap_{k=1}^{t-1}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_3(k))$$
$$= \Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k) \Big| \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t-1}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_3(k)) = 1,$$

which can be multiplied by

$$\Pr(\bigcap_{k=1}^{t-1}\Omega_1(k) \Big| \bigcap_{k=1}^{t-1}\Omega_2(k), \bigcap_{k=1}^{t-1}\Omega_4(k), \bigcap_{k=1}^{t-1}\Omega_5(k)).$$

The result is then multiplied by

$$\Pr(\bigcap_{k=1}^{t-1}\Omega_2(k)|\bigcap_{k=1}^{t-2}\Omega_1(k),\bigcap_{k=1}^{t-1}\Omega_3(k))=1.$$

Then the following equality holds:

$$\Pr(\bigcap_{k=1}^{t}\Omega_1(k),\bigcap_{k=1}^{t}\Omega_2(k)\Big|\bigcap_{k=1}^{t}\Omega_4(k),\bigcap_{k=1}^{t}\Omega_5(k),\bigcap_{k=1}^{t-2}\Omega_1(k),\bigcap_{k=1}^{t}\Omega_3(k))=1$$

which uses the fact that $\bigcap_{k=1}^{t}\Omega_y(k)\bigcap\bigcap_{k=1}^{t-1}\Omega_y(k)=\bigcap_{k=1}^{t}\Omega_y(k)$ $(y=1,2,3,4,5)$. So by repeating this until the iteration j_0, then the following equality holds:

$$\Pr(\bigcap_{k=1}^{t}\Omega_1(k),\bigcap_{k=1}^{t}\Omega_2(k)\Big|\bigcap_{k=1}^{t}\Omega_4(k),\bigcap_{k=1}^{t}\Omega_5(k),\bigcap_{k=1}^{j_0}\Omega_1(k),\bigcap_{k=1}^{t}\Omega_3(k))=1 \tag{58}$$

When $t\leq j_0$, we know that ${\mathbf{w}^{I,S}}_t={\mathbf{w}^{U,S}}_t$ and $M_1^S\geq M_1$, which means that if $\Omega_0(k)$ holds, then $\Omega_1(k)$ holds when ${\mathbf{w}^{I,S}}_t={\mathbf{w}^{U,S}}_t$, and thus

$$\Pr(\bigcap_{k=1}^{j_0}\Omega_1(k)|\bigcap_{k=1}^{j_0}\Omega_0(k))=1.$$

Then according to Theorem 8, we know that:

$$\Pr(\bigcap_{k=1}^{j_0}\Omega_0(k)\Big|\bigcap_{k=1}^{j_0}\Omega_5(k))=1.$$

By multiplying the above two formulas, we get:

$$\begin{aligned}&\Pr(\bigcap_{k=1}^{j_0}\Omega_1(k)|\bigcap_{k=1}^{j_0}\Omega_0(k))\cdot\Pr(\bigcap_{k=1}^{j_0}\Omega_0(k)\Big|\bigcap_{k=1}^{j_0}\Omega_5(k))\\&=\Pr(\bigcap_{k=1}^{j_0}\Omega_1(k),\bigcap_{k=1}^{j_0}\Omega_0(k)\Big|\bigcap_{k=1}^{j_0}\Omega_5(k))=1\end{aligned}$$

Note that since the probability of two joint events is smaller than that of either of the events, the following inequality holds:

$$\Pr(\bigcap_{k=1}^{j_0}\Omega_1(k),\bigcap_{k=1}^{j_0}\Omega_0(k)\Big|\bigcap_{k=1}^{j_0}\Omega_5(k))\leq\Pr(\bigcap_{k=1}^{j_0}\Omega_1(k)\Big|\bigcap_{k=1}^{j_0}\Omega_5(k))\leq 1.$$

So we know that:

$$\Pr(\bigcap_{k=1}^{j_0}\Omega_1(k)\Big|\bigcap_{k=1}^{j_0}\Omega_5(k))=1.$$

which can be multiplied by Formula (58) and thus the following equality holds:

$$\Pr(\bigcap_{k=1}^{t}\Omega_1(k),\bigcap_{k=1}^{t}\Omega_2(k)\Big|\bigcap_{k=1}^{t}\Omega_4(k),\bigcap_{k=1}^{t}\Omega_5(k),\bigcap_{k=1}^{t}\Omega_3(k))=1 \tag{59}$$

Then we can compute the probability of the negation of the joint event $(\bigcap_{k=1}^{t+1}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k))$:

$$\begin{aligned}
&\overline{\Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k))} \\
&= \overline{\Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k)) \Big| \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} \cdot \Pr(\bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k)) \\
&+ \Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k)) \Big| \overline{\bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} \cdot \overline{\Pr(\bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} \\
&\leq \overline{\Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k)) \Big| \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} + \overline{\Pr(\bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} \\
&= \overline{\Pr(\bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))}.
\end{aligned}$$

The last two steps use the fact that

$$\overline{\Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k)) \Big| \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} = 0$$

and

$$\overline{\Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k)) \Big| \overline{\bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} \leq 1.$$

By further using the property of the probability of the union of multiply events, the formula above is bounded as:

$$\begin{aligned}
\overline{\Pr(\bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k))} &\leq \Pr(\overline{\bigcap_{k=1}^{t}\Omega_4(k)} \bigcup \overline{\bigcap_{k=1}^{t}\Omega_5(k)} \bigcup \overline{\bigcap_{k=1}^{t}\Omega_3(k)}) \\
&\leq \Pr(\overline{\bigcap_{k=1}^{t}\Omega_4(k)}) + \Pr(\overline{\bigcap_{k=1}^{t}\Omega_5(k)}) + \Pr(\overline{\bigcap_{k=1}^{t}\Omega_3(k)}).
\end{aligned}$$

Then by using Theorem 6, Formula (53), Lemma 9 and taking the union between iteration 0 and t, we get:

$$\begin{aligned}
\overline{\Pr(\bigcap_{k=1}^{t}\Omega_3(k))} &\leq t \times \Psi_2, \\
\overline{\Pr(\bigcap_{k=1}^{t}\Omega_4(k))} &\leq t\Psi_3, \\
\overline{\Pr(\bigcap_{k=1}^{t}\Omega_5(k))} &\leq t \times \Psi_1.
\end{aligned}$$

Then we can know that:

$$\Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k)) \geq 1 - t(\Psi_2 + \Psi_3 + \Psi_1)$$

and thus

$$\Pr(\bigcap_{k=1}^{t}\Omega_1(k)) \geq \Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k)) \geq 1 - t(\Psi_2 + \Psi_3 + \Psi_1).$$

This finishes the proof.

Similarly, from Formula (59), we know that for all T iterations:

$$\Pr(\bigcap_{k=1}^{t}\Omega_1(k), \bigcap_{k=1}^{t}\Omega_2(k), \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k)\Big| \bigcap_{k=1}^{t}\Omega_4(k), \bigcap_{k=1}^{t}\Omega_5(k), \bigcap_{k=1}^{t}\Omega_3(k)) = 1. \tag{60}$$

Through the same argument, we know that:

$$\Pr(\bigcap_{k=1}^{T}\Omega_1(k), \bigcap_{k=1}^{T}\Omega_2(k), \bigcap_{k=1}^{T}\Omega_4(k), \bigcap_{k=1}^{T}\Omega_5(k)) \geq 1 - T(\Psi_2 + \Psi_3 + \Psi_1).$$

□

A.3.7 Proof of Theorem 11

The proof is the same as the proof of Theorem 4 except that $\mathbf{w}$, $\mathbf{w}^I$, n, r, ξ_{j_1,j_m}, $\mathbf{H}$, $\mathbf{B}$ need to be replaced by $\mathbf{w}^S$, $\mathbf{w}^{I,S}$, B, r, $\xi^S_{j_1,j_m}$, $\mathbf{H}^S$, $\mathbf{B}^S$ and the main theorems that the proof depends on will be replaced by Theorem 9 and Corollary 2. We will show some key steps below.

First of all, according to the proofs of Theorem 10, we know that the following inequalities hold with probability higher than $1 - T(\Psi_2 + \Psi_3 + \Psi_1)$:

$$\begin{aligned}
&\|{\mathbf{w}^S}_k - {\mathbf{w}^{I,S}}_k\| \leq \frac{1}{\frac{1}{2} - \frac{r}{n} - \frac{1}{B^{1/4}}} M_1(\frac{r}{n} + \frac{1}{B^{1/4}});\\
&\|{\mathbf{H}^S}_k - {\mathbf{B}^S}_{j_m}\| \leq \xi^S_{j_1,j_m};\\
&\xi^S_{j_0+xT_0,j_0+(x+m-1)T_0} \leq (1-\eta\mu)^{j_0+xT_0} A d_{0,mT_0-1} + A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}\\
&\leq \xi_{j_0,j_0+(m-1)T_0} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4};\\
&\frac{\Delta B_k}{B} \leq \frac{r}{n} + \frac{1}{B^{1/4}}.
\end{aligned}$$

Then by subtracting ${\mathbf{w}^{I,S}}_t$ by ${\mathbf{w}^{U,S}}_t$ and following the arguments from Formula (32) to (34), the following inequality holds for $\|{\mathbf{w}^{I,S}}_t - {\mathbf{w}^{U,S}}_t\|$ with probability higher than $1 - T \times (\Psi_2 + \Psi_1 + \Psi_3)$:

$$\begin{aligned}
&\|{\mathbf{w}^{I,S}}_t - {\mathbf{w}^{U,S}}_t\|\\
&\leq (1 - \eta\mu + \frac{B\eta}{B - \Delta B_t}\xi^S_{j_1,j_m})\|{\mathbf{w}^{I,S}}_t - {\mathbf{w}^{U,S}}_t\|\\
&+ \frac{\eta c_0}{2}\|{\mathbf{w}^{U,S}}_t - {\mathbf{w}^{I,S}}_t\|\|{\mathbf{w}^{U,S}}_t - {\mathbf{w}^{S}}_t\| + \frac{B\eta}{B - \Delta B_t}\xi^S_{j_1,j_m}\|{\mathbf{w}^{U,S}}_t - {\mathbf{w}^{S}}_t\|\\
&\leq \left(1 - \eta\mu + \frac{B\eta}{B - \Delta B_t}\xi^S_{j_1,j_m} + \frac{c_0 M_1\eta}{2}(\frac{r}{n} + \frac{1}{B^{1/4}})\right)\|{\mathbf{w}^{I,S}}_t - {\mathbf{w}^{U,S}}_t\| + \frac{B\eta}{B - \Delta B_t}\xi^S_{j_1,j_m} M_1(\frac{r}{n} + \frac{1}{B^{1/4}}).
\end{aligned}$$

By using the fact that $\frac{\Delta B_t}{B} \leq \frac{r}{n} + \frac{1}{B^{1/4}}$ and $\xi_{j_1,j_m} \leq \xi_{j_0,j_0+(m-1)T_0} + A \times M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}$, the formula

above can be bounded as:

$$
\begin{aligned}
&\|\mathbf{w}^{I,S}{}_t - \mathbf{w}^{U,S}{}_t\| \\
&\leq \left(1 - \eta\mu + \frac{B\eta}{B - \Delta B_t}\xi^S_{j_1,j_m} + \frac{c_0 M_1 \eta}{2}(\frac{r}{n} + \frac{1}{B^{1/4}})\right)\|\mathbf{w}^{I,S}{}_t - \mathbf{w}^{U,S}{}_t\| + \frac{B\eta}{B - \Delta B_t}\xi^S_{j_1,j_m} M_1(\frac{r}{n} + \frac{1}{B^{1/4}}) \\
&\leq [1 - \eta\mu + \frac{\eta}{1 - \frac{r}{n} - \frac{1}{B^{1/4}}}(\xi^S_{j_0,j_0+(m-1)T_0} + A \times M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}) \\
&+ \frac{c_0 M_1 \eta}{2}(\frac{r}{n} + \frac{1}{B^{1/4}})]\|\mathbf{w}^{I,S}{}_t - \mathbf{w}^{U,S}{}_t\| + \frac{\eta}{1 - \frac{r}{n} - \frac{1}{B^{1/4}}}\xi^S_{j_0+xT_0,j_0+(x+m-1)T_0} M_1(\frac{r}{n} + \frac{1}{B^{1/4}}).
\end{aligned}
$$

Since $\xi^S_{j_0,j_0+(m-1)T_0} + A \times M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4} \leq \frac{\mu}{2}$ and B is a large mini-batch size, then

$$
1 - (\eta\mu - \frac{\eta}{1 - \frac{r}{n} - \frac{1}{B^{1/4}}}(\xi^S_{j_0,j_0+(m-1)T_0} + A \times M_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}) - \frac{c_0 M_1 \eta}{2}(\frac{r}{n} + \frac{1}{B^{1/4}})) < 1.
$$

Then after explicitly using the definition of $\xi^S_{j_1,j_m}$ and following the argument of equation (35) to (38), we get:

$$
\begin{aligned}
&\|\mathbf{w}^I{}_{j_0+(y+m)T_0} - \mathbf{w}^U{}_{j_0+(y+m)T_0}\| \\
&\leq (1 - \eta C)^{yT_0}\|\mathbf{w}^I{}_{j_0+mT_0} - \mathbf{w}^U{}_{j_0+mT_0}\| \\
&+ \frac{M_1(\frac{r}{n} + \frac{1}{B^{1/4}})}{C(1 - \frac{r}{n} - \frac{1}{B^{1/4}})}(1 - \eta C)^{yT_0}(1 - \eta\mu)^{j_0} d_{0,mT_0-1}\frac{1}{1 - (\frac{1-\eta\mu}{1-\eta C})^{T_0}} \\
&+ \frac{1}{1 - (1 - \eta C)^{T_0}}\frac{M_1(\frac{r}{n} + \frac{1}{B^{1/4}})}{C(1 - \frac{r}{n} - \frac{1}{B^{1/4}})}(A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4})
\end{aligned} \tag{61}
$$

when $t \to \infty$ and thus $y \to \infty$, $(1 - \eta C)^{yT_0} \to 0$. Also with large mini-batch value B, $A_1\frac{r}{n} + A_2\frac{1}{B^{1/4}} + AM_1\left(\frac{(\log(p+1))^2}{B}\right)^{1/4}$ is a value of the same order as $\frac{r}{n} + \frac{1}{B^{1/4}}$. Thus

$$
\|\mathbf{w}^I{}_{j_0+(y+m)T_0} - \mathbf{w}^U{}_{j_0+(y+m)T_0}\| = o(\frac{r}{n} + \frac{1}{B^{1/4}})
$$

and

$$
\|\mathbf{w}^{U,S}{}_t - \mathbf{w}^{I,S}{}_t\| \leq o(\frac{r}{n} + \frac{1}{B^{1/4}}).
$$

B Details on applications

B.1 Privacy related data deletion

The notion of Approximate Data Deletion from the training dataset is proposed in Ginart et al. (2019):

Definition 1. *A data deletion operation R_A is a $\delta-$deletion for algorithm A if, for all datasets D and for all measurable subset S, the following inequality holds:*

$$
Pr[A(D_{-i}) \in S|D_{-i}] \geq \delta Pr[R_A(D, A(D), i) \in S|D_{-i}],
$$

where D is the full training dataset, D_{-i} is the remaining dataset after the i_{th} sample is removed, $A(D)$ and $A(D_{-i})$ represent the model trained over D and D_{-i} respectively. Also R_A is an approximate model update algorithm, which updates the model after the sample i is removed.

This definition mimics the classical definition of differential privacy (Dwork et al., 2014):

Definition 2. *A mechanism M is ϵ-differentially private, where $\epsilon \geq 0$, if for all neighboring databases D_0 and D_1, i.e., for databases differing in only one record, and for all sets $S \in [M]$, where $[M]$ is the range of M, the following inequality holds:*

$$Pr[M(D_0) \in S] \leq e^{\epsilon} Pr[M(D_1) \in S].$$

By borrowing the notations from Ginart et al. (2019), we define a version of approximate data deletion, which is slightly more strict than the one from Ginart et al. (2019):

Definition 3. *R_A is an $\epsilon-$approximate deletion for A if for all D and measurable subset $S \subset \mathcal{H}$:*

$$P(A(D_{-i}) \in S|D_{-i}) \leq e^{\epsilon} P(R_A(D, A(D), i) \in S|D_{-i})$$

and

$$P(R_A(D, A(D), i) \in S|D_{-i}) \leq e^{\epsilon} P(A(D_{-i}) \in S|D_{-i}).$$

To satisfy this definition for gradient descent, necessary randomness is added to the output of the BaseL and DeltaGrad. One simple way is the Laplace mechanism (Dwork et al., 2014), also following the idea from Chaudhuri and Monteleoni (2009) where noise following the Laplace distribution, i.e.

$$Lap(x|\frac{2}{n\epsilon\lambda}) = \frac{1}{\frac{2}{n\epsilon\lambda}} \exp(-\frac{|x|}{\frac{2}{n\epsilon\lambda}}),$$

is added to the each coordinate of the output of the regularized logistic regression. Here p is the number of the parameters, λ is the regularization rate and $\frac{2}{n\lambda}$ is the *sensitivity* of logistic regression (see Chaudhuri and Monteleoni (2009) for more details).

We can add even smaller noise to $\mathbf{w}^*$, $\mathbf{w}^{U^*}$ and $\mathbf{w}^{I^*}$, which follows the distribution $Lap(\frac{\delta}{\epsilon})$ for each coordinate of $\mathbf{w}^*$, $\mathbf{w}^{U^*}$ and $\mathbf{w}^{I^*}$ and is independent across different coordinates. Here $\delta > \sqrt{p}\delta_0$ and

$$\delta_0 = \frac{1}{\eta(\frac{1}{2}\mu - \frac{r}{n-r}\mu - \frac{c_0 M_1 r}{2n})^2} \frac{M_1 r}{n-r}(A\frac{1}{\frac{1}{2} - \frac{r}{n}} M_1 \frac{r}{n})$$

(which is an upper bound on $\|\mathbf{w}^{U^*} - \mathbf{w}^{I^*}\|$), such that the randomized DeltaGrad preserves $\epsilon-$approximate deletion.

Proof. We denote the model parameters after adding the random noise over $\mathbf{w}^R$, $\mathbf{w}^{U,R}$ and $\mathbf{w}^{I,R}$, and $\mathbf{v}_i$ as the value of $\mathbf{v}$ in the i_{th} coordinate. We have:

$$\mathbf{w}^* - \mathbf{w}^{R^*}, \mathbf{w}^{U^*} - \mathbf{w}^{U,R^*}, \mathbf{w}^{I^*} - \mathbf{w}^{I,R^*} \sim Lap(\frac{\delta}{\epsilon})$$

Given an arbitrary vector $\mathbf{z} = [z_1, z_2, \ldots, z_p]$, the probability density ratio between $Pdf(\mathbf{w}^{U,R^*} = \mathbf{z})$ and $Pdf(\mathbf{w}^{I,R^*} = \mathbf{z})$ can be calculated as

$$\begin{aligned}
&\frac{Pdf(\mathbf{w}^{U,R^*} = \mathbf{z})}{Pdf(\mathbf{w}^{I,R^*} = \mathbf{z})} = \frac{\Pi_{i=1}^p \frac{\epsilon}{\delta} \exp(-\frac{\epsilon|\mathbf{Z} - \mathbf{w}^{U^*}|}{\delta})}{\Pi_{i=1}^p \frac{\epsilon}{\delta} \exp(-\epsilon\frac{|\mathbf{z}_i - \mathbf{w}^{I^*}_i|}{\delta})} \\
&= \Pi_{i=1}^p \exp(\frac{\epsilon(|\mathbf{z}_i - \mathbf{w}^{U^*}| - |\mathbf{z}_i - \mathbf{w}^{I^*}_i|)}{\delta}) \\
&\leq \Pi_{i=1}^p \exp(\frac{\epsilon(|\mathbf{w}^{I^*}_i - \mathbf{w}^{U^*}_i|)}{\delta}) \\
&= \exp(\frac{\epsilon(\|\mathbf{w}^{I^*} - \mathbf{w}^{U^*}\|_1)}{\delta})
\end{aligned}$$

Since

$$\|\mathbf{w}^{I^*} - \mathbf{w}^{U^*}\|_1 \leq \sqrt{p}\|\mathbf{w}^{I^*} - \mathbf{w}^{U^*}\|_2 = \sqrt{p}\|\mathbf{w}^{I^*} - \mathbf{w}^{U^*}\|$$

Then,

$$\frac{Pdf(\mathbf{w}^{U,R^*}=\mathbf{z})}{Pdf(\mathbf{w}^{I,R^*}=\mathbf{z})} \leq \exp(\frac{\epsilon(\|\mathbf{w}^{I^*}-\mathbf{w}^{U^*}\|)}{\delta})$$
$$\leq \exp(\frac{\epsilon\sqrt{p}\delta_0}{\delta}) \leq \exp(\epsilon)$$

Similarly, we can also prove $\frac{Pdf(\mathbf{w}^{U,R^*}=\mathbf{z})}{Pdf(\mathbf{w}^{I,R^*}=\mathbf{z})} \geq \exp(\epsilon)$ by symmetry.

□

C Supplementary algorithm details

In Section 2, we only provided the details of DeltaGrad for deterministic gradient descent for the strongly convex and smooth objective functions in batch deletion/addition scenarios. In this section, we will provide more details on how to extend DeltaGrad to handle stochastic gradient descent, online deletion/addition scenarios and non-strongly convex, non-smooth objective functions.

C.1 Extension of DeltaGrad for stochastic gradient descent

By using the notations from equations (5)-(7), we need to approximately or explicitly compute $G_{B,S}$, i.e. the average gradient for a mini-batch in the SGD version of DeltaGrad, instead of ∇F , which is the average gradient for all samples. So by replacing $\mathbf{w}_t$, $\mathbf{w}^U{}_t$, $\mathbf{w}^I{}_t$, ∇F, $\mathbf{B}$ and $\mathbf{H}$ with $\mathbf{w}^S{}_t$, $\mathbf{w}^{U,S}{}_t$, $\mathbf{w}^{I,S}{}_t$, $G_{B,S}$, $\mathbf{B}^S$ and $\mathbf{H}^S$ in Algorithm 1, we get the SGD version of DeltaGrad.

C.2 Extension of DeltaGrad for online deletion/addition

In the online deletion/addition scenario, whenever the model parameters are updated after the deletion or addition of one sample, the history information should be also updated to reflect the changes. By assuming that only one sample is deleted or added each time, the online deletion/addition version of DeltaGrad is provided in Algorithm 2 and the differences relative to Algorithm 1 are highlighted.

Since the history information needs to be updated every time when new deletion or addition requests arrive, we need to do some more analysis on the error bound, which is still pretty close to the analysis in Section A.

In what follows, the analysis will be conducted on gradient descent with online deletion. Other similar scenarios, e.g. stochastic gradient descent with online addition, will be left as the future work.

C.2.1 Convergence rate analysis for online gradient descent version of DeltaGrad

Additional notes on setup, preliminaries

Let us still denote the model parameters for the original dataset at the t_{th} iteration by $\mathbf{w}_t$. During the model update phase for the k_{th} deletion request at the t_{th} iteration, the model parameters updated by BaseL and DeltaGrad are denoted by $\mathbf{w}^U{}_t(k)$ and $\mathbf{w}^I{}_t(k)$ respectively where $\mathbf{w}^U{}_t(0) = \mathbf{w}^I{}_t(0) = \mathbf{w}_t$. We also assume that the total number of removed samples in all deletion requests, r, is still far smaller than the total number of samples, n.

Also suppose that the indices of the removed samples are $\{i_1, i_2, \ldots, i_r\}$, which are removed at the 1_{st}, 2_{nd}, $3_{rd}, \ldots,$ r_{th} deletion request. This also means that the cumulative number of samples up to the k_{th} deletion request ($k \leq r$) is $n-k$ for all $1 \leq k \leq r$ and thus the objective function at the k_{th} iteration will be:

$$F^k(\mathbf{w}) = \frac{1}{n-k}\sum_{i\notin R_k} F_i(\mathbf{w}).$$

where $R_k = \{i_1, i_2, \ldots, i_k\}$. Plus, at the k_{th} deletion request, we denote by $\mathbf{H}_t^k$ the average Hessian matrix of $F^k(\mathbf{w})$ evaluated between $\mathbf{w}^I{}_t(k+1)$ and $\mathbf{w}^I{}_t(k)$:

$$\mathbf{H}_t^k = \frac{1}{n-k}\sum_{i\notin R_k}\int_0^1 \mathbf{H}_i(\mathbf{w}^I{}_t(k) + x(\mathbf{w}^I{}_t(k+1) - \mathbf{w}^I{}_t(k)))dx$$

Algorithm 2: DeltaGrad (online deletion/addition)

Input : The full training set $(\mathbf{X}, \mathbf{Y})$, model parameters cached during the training phase for the full training samples $\{\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t\}$ and corresponding gradients $\{\nabla F(\mathbf{w}_0), \nabla F(\mathbf{w}_1), \ldots, \nabla F(\mathbf{w}_t)\}$, the index of the removed training sample or the added training sample i_r, period T_0, total iteration number T, history size m, warmup iteration number j_0, learning rate η

Output: Updated model parameter ${\mathbf{w}^I}_t$

Initialize ${\mathbf{w}^I}_0 \leftarrow \mathbf{w}_0$
Initialize an array $\Delta G = []$
Initialize an array $\Delta W = []$
for $t = 0; t < T; t++$ **do**
  if $[((t - j_0) \mod T_0) == 0]$ *or* $t \leq j_0$ **then**
    compute $\nabla F\left({\mathbf{w}^I}_t\right)$ exactly
    compute $\nabla F\left({\mathbf{w}^I}_t\right) - \nabla F(\mathbf{w}_t)$ based on the cached gradient $\nabla F(\mathbf{w}_t)$
    set $\Delta G[k] = \nabla F\left({\mathbf{w}^I}_t\right) - \nabla F(\mathbf{w}_t)$
    set $\Delta W[k] = {\mathbf{w}^I}_t - \mathbf{w}_t$, based on the cached parameters $\mathbf{w}_t$
    $k \leftarrow k + 1$
    compute ${\mathbf{w}^I}_{t+1}$ by using exact GD update (equation (1))
    $\mathbf{w}_t \leftarrow {\mathbf{w}^I}_t$
    $\nabla F(\mathbf{w}_t) \leftarrow \nabla F({\mathbf{w}^I}_t)$
  else
    Pass $\Delta W[-m:]$, $\Delta G[-m:]$, the last m elements in ΔW and ΔG, which are from the $j_1^{th}, j_2^{th}, \ldots, j_m^{th}$ iterations where $j_1 < j_2 < \cdots < j_m$ depend on t, $\mathbf{v} = {\mathbf{w}^I}_t - \mathbf{w}_t$, and the history size m, to the L-BFGFS Algorithm (See Supplement) to get the approximation of $\mathbf{H}(\mathbf{w}_t)\mathbf{v}$, i.e., $\mathbf{B}_{j_m}\mathbf{v}$
    Approximate $\nabla F\left({\mathbf{w}^I}_t\right) = \nabla F(\mathbf{w}_t) + \mathbf{B}_{j_m}\left({\mathbf{w}^I}_t - \mathbf{w}_t\right)$
    Compute ${\mathbf{w}^I}_{t+1}$ by using the "leave-1-out" gradient formula, based on the approximated $\nabla F({\mathbf{w}^I}_t)$
    $\mathbf{w}_t \leftarrow {\mathbf{w}^I}_t$
    $\nabla F(\mathbf{w}_t) \leftarrow \frac{\eta}{n-1}[n(\mathbf{B}_{j_m}({\mathbf{w}^I}_t - \mathbf{w}_t) + \nabla F(\mathbf{w}_t)) - \nabla F_{i_r}(\mathbf{w}_t)]$
  end
end
return ${\mathbf{w}^I}_t$

Specifically,

$$\mathbf{H}_t^0 = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathbf{H}_i({\mathbf{w}^I}_t(0) + x({\mathbf{w}^I}_t(1) - {\mathbf{w}^I}_t(0)))dx.$$

Also the model parameters and the approximate gradients evaluated by DeltaGrad at the $r-1_{st}$ deletion request are used at the r_{th} request, and are denoted by:

$$\{{\mathbf{w}^I}_0(r-1), {\mathbf{w}^I}_1(r-1), \ldots, {\mathbf{w}^I}_t(r-1)\}$$

and

$$\{g^a\left({\mathbf{w}^I}_0(r-1)\right), g^a\left({\mathbf{w}^I}_1(r-1)\right), \ldots, g^a\left({\mathbf{w}^I}_t(r-1)\right)\}.$$

Note that $g^a({\mathbf{w}^I}_t(k))$ $(k \leq r)$ is not necessarily equal to ∇F due to the approximation brought by DeltaGrad. But due to the periodicity of DeltaGrad, at iteration $0, 1, \ldots, j_0$ and iteration $j_0 + xT_0$ $(x = 1, 2, \ldots,)$, the gradients are explicitly evaluated, i.e.:

$$g^a({\mathbf{w}^I}_t(k)) = \frac{1}{n-k}\sum_{i \notin R_k} \nabla F_i({\mathbf{w}^I}_t(k))$$

for $t = 0, 1, \ldots, j_0$ or $t = j_0 + xT_0$ $(x \geq 1)$ and all $k \leq r$.

Also, due to the periodicity, the sequence $[\Delta g_{j_0}, \Delta g_{j_1}, \ldots, \Delta g_{j_{m-1}}]$ used in approximating the Hessian matrix always uses the exact gradient information, which means that:

$$\Delta g_{j_q} = \frac{1}{n-k}[\sum_{i \notin R_k} \nabla F_i({\mathbf{w}^I}_{j_q}(k)) - \sum_{i \notin R_k} \nabla F_i({\mathbf{w}^I}_{j_q}(k-1))]$$

where $q = 1, 2, \ldots, m-1$. So Lemma 6 on the bound on the eigenvalues of B_{j_q} holds for all q and $k = 1, 2, \ldots, r$.

But for the iterations where the gradients are not explicitly evaluated, the calculation of $g^a(\mathbf{w}^I{}_t(k))$ depends on the approximated Hessian matrix $\mathbf{B}_{j_m}^{k-1}$ and the approximated gradients calculated at the t_{th} iteration at the $k-1$-st deletion request. So the update rule for $g^a(\mathbf{w}^I{}_t(k))$ is:

$$\begin{aligned} g^a(\mathbf{w}^I{}_t(k)) = \frac{1}{n-k}\{(n-k+1)[\mathbf{B}_{j_m}^{k-1}(\mathbf{w}^I{}_t(k) - \mathbf{w}^I{}_t(k-1)) \\ + g^a(\mathbf{w}^I{}_t(k-1))] - \nabla F_{i_k}(\mathbf{w}^I{}_t(k))\}. \end{aligned} \tag{62}$$

Here the product $\mathbf{B}_{j_m}^{k-1} \cdot (\mathbf{w}^I{}_t(k) - \mathbf{w}^I{}_t(k-1))$ approximates

$$\frac{1}{n-k+1} \sum_{i \notin R_{k-1}} \nabla F_i(\mathbf{w}^I{}_t(k)) - \nabla F_i(\mathbf{w}^I{}_t(k-1))$$

and $g^a(\mathbf{w}^I{}_t(k-1))$ approximates $\frac{1}{n-k+1}\sum_{i \notin R_{k-1}} \nabla F_i(\mathbf{w}^I{}_t(k-1))$.

Similarly, the online version of Δw (at the k_{th} iteration) becomes:

$$\Delta w_{j_q}(k) = \mathbf{w}^I{}_{j_q}(k) - \mathbf{w}^I{}_{j_q}(k-1)$$

where $q = 1, 2, \ldots, m-1$.

Similarly, we use $d_{j_a, j_b}(k)$ to denote the value of the upper bound d on the distance between the iterates at the k_{th} deletion request and use $\mathbf{B}_{j_m}^{k-1}$ to denote the approximated Hessian matrix in the k_{th} deletion request, which approximated the Hessian matrix $\mathbf{H}_t^{k-1}$.

So the update rule for $\mathbf{w}^I{}_t(k)$ becomes:

$$\mathbf{w}^I{}_{t+1}(k) = \begin{cases} \mathbf{w}^I{}_t(k) - \frac{\eta}{n-k}\sum_{i \notin R_k} \nabla F_i(\mathbf{w}^I{}_t(k)),\ [(t-j_0) \mod T_0 = 0] \text{ or } t \leq j_0 \\ \mathbf{w}^I{}_t(k) - \frac{\eta}{n-k}\{(n-k+1)[\mathbf{B}_{j_m}^{k-1}(\mathbf{w}^I{}_t(k) - \mathbf{w}^I{}_t(k-1)) \\ \quad + g^a(\mathbf{w}^I{}_t(k-1))] - \nabla F_{i_k}(\mathbf{w}^I{}_t(k))\},\ else. \end{cases} \tag{63}$$

Proof preliminaries.

On each deletion request, the BaseL model parameters are retrained from scratch on the remaining samples. This implies that Theorem 2 still holds, if we replace $\mathbf{w}^U{}_t$, $\mathbf{w}_t$ and r with $\mathbf{w}^U{}_t(k)$, $\mathbf{w}^U{}_t(k-1)$ and 1 respectively:

Theorem 12 (Bound between iterates deleting one datapoint)**.** $\|\boldsymbol{w}^U{}_t(r) - \boldsymbol{w}^U{}_t(r-1)\| \leq M_1 \frac{1}{n}$ *where* $M_1 = \frac{2}{\mu}c_2$ *is some positive constant that does not depend on* t*. Here* μ *is the strong convexity constant, and* c_2 *is the bound on the individual gradients.*

By induction, we have:

$$\|\mathbf{w}^U{}_t(r) - \mathbf{w}_t\| = \|\mathbf{w}^U{}_t(r) - \mathbf{w}^U{}_t(0)\| \leq M_1 \frac{r}{n}. \tag{64}$$

Then let us do some analysis on $d_{j_a, j_b}(k)$. We use the notation $M_1^r \frac{1}{n}$ for $\frac{\frac{2M_1}{n}}{1 - \frac{r+1}{n} - \frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})}$, where M_1^r is a constant which does not depend on k.

Lemma 11. *If* $\|\boldsymbol{w}^I{}_t(k) - \boldsymbol{w}^I{}_t(k-1)\| \leq \frac{\frac{2M_1}{n}}{1 - \frac{k+1}{n} - \frac{2(k-1)}{n}(\frac{2L+\mu}{\mu})}$ *for all* $k \leq r$*, then* $d_{j_a, j_b}(r) \leq d_{j_a, j_b}(0) + 2r \cdot M_1^r \frac{1}{n}$ *where* M_1 *is defined in Theorem 12.*

Proof. Recall that $d_{j_a, j_b}(k) = \max(\|\mathbf{w}^I{}_y(k) - \mathbf{w}^I{}_z(k)\|)_{j_a < y < z < j_b}$. Then for two arbitrary iterations y, z, let us bound $\|\mathbf{w}^I{}_y(k) - \mathbf{w}^I{}_z(k)\|$ as below:

$$\begin{aligned} &\|\mathbf{w}^I{}_y(k) - \mathbf{w}^I{}_z(k)\| \\ &= \|\mathbf{w}^I{}_y(k) - \mathbf{w}^I{}_z(k) + \mathbf{w}^I{}_y(k-1) - \mathbf{w}^I{}_z(k-1) + \mathbf{w}^I{}_z(k-1) - \mathbf{w}^I{}_y(k-1)\| \\ &\leq \|\mathbf{w}^I{}_y(k) - \mathbf{w}^I{}_y(k-1)\| + \|\mathbf{w}^I{}_z(k) - \mathbf{w}^I{}_z(k-1)\| + \|\mathbf{w}^I{}_z(k-1) - \mathbf{w}^I{}_y(k-1)\|. \end{aligned}$$

Then by using the bound on $\|\mathbf{w}^I{}_t(k) - \mathbf{w}^I{}_t(k-1)\|$, the above formula leads to:

$$\leq 2 \cdot \frac{\frac{2M_1}{n}}{1 - \frac{k+1}{n} - \frac{2(k-1)}{n}(\frac{2L+\mu}{\mu})} + \|\mathbf{w}^I{}_z(k-1) - \mathbf{w}^I{}_y(k-1)\|.$$

By using that $\frac{\frac{2M_1}{n}}{1-\frac{k+1}{n}-\frac{2(k-1)}{n}(\frac{2L+\mu}{\mu})} \leq \frac{\frac{2M_1}{n}}{1-\frac{r+1}{n}-\frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})}$ and applying it recursively for $k = 1, 2, \ldots, r$, we have:

$$\|\mathbf{w}^I{}_y(r) - \mathbf{w}^I{}_z(r)\| \leq 2r \cdot \frac{\frac{2M_1}{n}}{1 - \frac{r+1}{n} - \frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})} + \|\mathbf{w}^I{}_z(0) - \mathbf{w}^I{}_y(0)\|.$$

Then by using the definition of $d_{j_a,j_b}(k)$, the following inequality holds:

$$d_{j_a,j_b}(r) \leq d_{j_a,j_b+T_0-1}(0) + 2r \cdot \frac{\frac{2M_1}{n}}{1 - \frac{r+1}{n} - \frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})}.$$

Recalling the definition of $M_1^r \frac{1}{n}$, this is exactly the required result. □

We also mention that, since $\frac{\frac{2M_1}{n}}{1-\frac{k+1}{n}-\frac{2(k-1)}{n}(\frac{2L+\mu}{\mu})} \leq \frac{\frac{2M_1}{n}}{1-\frac{r+1}{n}-\frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})}$, then $\|\mathbf{w}^I{}_t(k) - \mathbf{w}^I{}_t(k-1)\| \leq M_1^r \frac{1}{n}$ for any $k \leq r$.

Theorem 13. *Suppose that at the k_{th} deletion request, $\|\boldsymbol{w}^I{}_{j_q}(k) - \boldsymbol{w}^I{}_{j_q}(k-1)\| \leq M_1^r \frac{1}{n}$, where $q = 1, 2, \ldots, m$ and $M_1 = \frac{2c_2}{\mu}$. Let $e = \frac{L(L+1)+K_2L}{\mu K_1}$ for the upper and lower bounds K_1, K_2 on the eigenvalues of the quasi-Hessian from Lemma 6, and for the Lipshitz constant c_0 of the Hessian. For $1 \leq z+1 \leq y \leq m$ we have:*

$$\|\boldsymbol{H}_{j_z}^{k-1} - \boldsymbol{H}_{j_y}^{k-1}\| \leq c_0 d_{j_z,j_y}(k-1) + c_0 M_1^r \frac{1}{n}$$

and

$$\|\Delta g_{j_z} - \boldsymbol{B}_{j_y}^{k-1} \Delta w_{j_z}\| \leq \left[(1+e)^{y-z-1} - 1\right] \cdot c_0 (d_{j_z,j_y} + M_1^r \frac{1}{n}) \cdot s_{j_1,j_m}(k-1)$$

where $s_{j_1,j_m}(k-1) = \max\left(\|\Delta w_a(k-1)\|\right)_{a=j_1,j_2,\ldots,j_m} = \max\left(\|\boldsymbol{w}^{I,S}{}_a(k-1) - \boldsymbol{w}^{I,S}{}_a(k-2)\|\right)_{a=j_1,j_2,\ldots,j_m}$. Recall that d is defined as the maximum gap between the steps of the algorithm for the iterations from j_z to j_y:

$$d_{j_z,j_y}(k-1) = \max\left(\|\boldsymbol{w}^I{}_a(k-1) - \boldsymbol{w}^I{}_b(k-1)\|\right)_{j_z \leq a \leq b \leq j_y}. \tag{65}$$

Proof. Let us bound the difference between the averaged Hessians $\|\mathbf{H}_{j_z}^{k-1} - \mathbf{H}_{j_y}^{k-1}\|$, where $1 \leq z < y \leq m$, using their definition, as well as using Assumption 4 on the Lipshitzness of the Hessian. First we can get the following equality:

$$\begin{aligned}
&\|\mathbf{H}_{j_y}^{k-1} - \mathbf{H}_{j_z}^{k-1}\| \\
&= \|\int_0^1 [\mathbf{H}(\mathbf{w}^I{}_{j_y}(k-1) + x(\mathbf{w}^I{}_{j_y}(k) - \mathbf{w}^I{}_{j_y}(k-1)))]dx \\
&- \int_0^1 [\mathbf{H}(\mathbf{w}^I{}_{j_z}(k-1) + x(\mathbf{w}^I{}_{j_z}(k) - \mathbf{w}^I{}_{j_z}(k-1)))]dx\| \\
&= \|\int_0^1 [\mathbf{H}(\mathbf{w}^I{}_{j_y}(k-1) + x(\mathbf{w}^I{}_{j_y}(k) - \mathbf{w}^I{}_{j_y}(k-1))) \\
&- \mathbf{H}(\mathbf{w}^I{}_{j_z}(k-1) + x(\mathbf{w}^I{}_{j_z}(k) - \mathbf{w}^I{}_{j_z}(k-1)))]dx\|
\end{aligned} \tag{66}$$

Then we can bound this as:

$$
\begin{aligned}
&\leq c_0 \int_0^1 \|\mathbf{w}^I{}_{j_y}(k-1) + x(\mathbf{w}^I{}_{j_y}(k) - \mathbf{w}^I{}_{j_y}(k-1)) \\
&- [\mathbf{w}^I{}_{j_z}(k-1) + x(\mathbf{w}^I{}_{j_z}(k) - \mathbf{w}^I{}_{j_z}(k-1))]\|dx \\
&\leq c_0 \|\mathbf{w}^I{}_{j_y}(k-1) - \mathbf{w}^I{}_{j_z}(k-1)\| \\
&+ \frac{c_0}{2}\|\mathbf{w}^I{}_{j_y}(k) - \mathbf{w}^I{}_{j_y}(k-1) - (\mathbf{w}^I{}_{j_z}(k) - \mathbf{w}^I{}_{j_z}(k-1))\| \\
&\leq c_0 \|\mathbf{w}^I{}_{j_y}(k-1) - \mathbf{w}^I{}_{j_z}(k-1)\| \\
&+ \frac{c_0}{2}\|\mathbf{w}^I{}_{j_z}(k) - \mathbf{w}^I{}_{j_z}(k-1)\| + \frac{c_0}{2}\|\mathbf{w}^I{}_{j_y}(k) - \mathbf{w}^I{}_{j_y}(k-1)\| \\
&\leq c_0 d_{j_y,j_z}(k-1) + c_0 M_1^r \frac{1}{n} \leq c_0 d_{j_1,j_m+T_0-1}(k-1) + c_0 M_1^r \frac{1}{n}.
\end{aligned}
$$

On the last line, we have used the definition of d_{j_z,j_y}, and the assumption on the boundedness of $\|\mathbf{w}^I{}_{j_z}(k) - \mathbf{w}^I{}_{j_z}(k-1)\|$.

Then by following the rest of the proof of Theorem 1, we get:

$$
\|\Delta g_{j_z} - \mathbf{B}_{j_y} \Delta w_{j_z}\| \leq \left[(1+e)^{y-z-1} - 1\right] \cdot c_0 (d_{j_z,j_y}(k-1) + M_1^r \frac{1}{n}) \cdot s_{j_1,j_m}(k-1).
$$

□

Similarly, the online version of Corollary 1 also holds by following the same derivation as the proof of Corollary 1 (except that r, ξ_{j_1,j_m} and d_{j_1,j_m+T_0-1} is replaced by 1, $\xi_{j_1,j_m}(k-1)$ and $d_{j_1,j_m+T_0-1}(k-1)$ respectively), i.e.:

Corollary 3 (Approximation accuracy of quasi-Hessian to mean Hessian (online deletion))**.** *Suppose that at the k_{th} deletion request, $\|\boldsymbol{w}^I{}_{j_s}(k) - \boldsymbol{w}^I{}_{j_s}(k-1)\| \leq M_1^r \frac{1}{n}$ and $\|\boldsymbol{w}^I{}_t(k) - \boldsymbol{w}^I{}_t(k-1)\| \leq M_1^r \frac{1}{n}$ where $s = 1, 2, \ldots, m$. Then for $j_m \leq t \leq j_m + T_0 - 1$,*

$$
\|\boldsymbol{H}_t^{k-1} - \boldsymbol{B}_{j_m}^{k-1}\| \leq \xi_{j_1,j_m}(k-1) := A d_{j_1,j_m+T_0-1}(k-1) + A M_1^r \frac{1}{n}. \tag{67}
$$

Recall that $A = \frac{c_0\sqrt{m}[(1+e)^m - 1]}{c_1} + c_0$, where c_0 is the Lipschitz constant of the Hessian, c_1 is the "strong independence" constant from Assumption 5, and $d_{j_1,j_m+T_0-1}(k-1)$ is the maximal gap between the iterates of the GD algorithm on the full data from j_1 to $j_m + T_0 - 1$ after the $k-1$-st deletion.

Based on this, let us derive a bound on $\|\nabla F_i(\mathbf{w}^I{}_t(r))\|$, $\|g^a(\mathbf{w}^I{}_t(r)) - \frac{1}{n-r}\sum_{i \notin R_r} \nabla F_i(\mathbf{w}^I{}_t(r))\|$ and $\|g^a(\mathbf{w}^I{}_t(r))\|$.

Lemma 12. *Suppose we are at an iteration t such that $j_m \leq t \leq j_m + T_0 - 1$. If the following inequality holds for all $k < r$:*

$$
\|\boldsymbol{w}^I{}_t(k) - \boldsymbol{w}^I{}_t(k-1)\| \leq M_1^r \frac{1}{n},
$$

then the following inequality holds for all $i = 1, 2, \ldots, n$:

$$
\|\nabla F_i(\boldsymbol{w}^I{}_t(r-1))\| \leq M_1^r \frac{1}{n} L r + c_2.
$$

Proof. By adding and subtracting $\nabla F_i(\mathbf{w}^I{}_t(r-2))$ inside $\|\nabla F_i(\mathbf{w}^I{}_t(r-1))\|$, we get:

$$
\begin{aligned}
&\|\nabla F_i(\mathbf{w}^I{}_t(r-1))\| \\
&= \|\nabla F_i(\mathbf{w}^I{}_t(r-1)) - \nabla F_i(\mathbf{w}^I{}_t(r-2)) + \nabla F_i(\mathbf{w}^I{}_t(r-2))\| \\
&\leq \|\nabla F_i(\mathbf{w}^I{}_t(r-1)) - \nabla F_i(\mathbf{w}^I{}_t(r-2))\| + \|\nabla F_i(\mathbf{w}^I{}_t(r-2))\|
\end{aligned}
$$

The last inequality uses the triangle inequality. Then by using the Cauchy mean value theorem, the upper bound on the eigenvalue of the Hessian matrix (i.e. Assumption 2) and the bound on $\|\mathbf{w}^I{}_t(k)-\mathbf{w}^I{}_t(k-1)\|$, the formula above is bounded as (recall $\mathbf{H}_i$ is an integrated Hessian):

$$= \|\mathbf{H}_i(\mathbf{w}^I{}_t(r-1)+x(\mathbf{w}^I{}_t(r-2)-\mathbf{w}^I{}_t(r-1)))\cdot(\mathbf{w}^I{}_t(r-1)-\mathbf{w}^I{}_t(r-2))\|+\|\nabla F_i(\mathbf{w}^I{}_t(r-2))\|$$
$$\leq LM_1^r\frac{1}{n}+\|\nabla F_i(\mathbf{w}^I{}_t(r-2))\|.$$

By using this recursively, we get:

$$\leq \sum_{k=1}^{r-1} M_1^r\frac{1}{n}L+\|\nabla F_i(\mathbf{w}^I{}_t(0))\|\leq M_1^r\frac{1}{n}Lr+c_2.$$

□

Lemma 13. *If at a given iteration t such that $j_m \leq t \leq j_m+T_0-1$, for all $k<r$, the following inequalities hold:*

$$\|\boldsymbol{w}^I{}_t(k)-\boldsymbol{w}^I{}_t(k-1)\|\leq M_1^r\frac{1}{n}$$

and

$$\xi_{j_1,j_m}(k-1)\leq\frac{\mu}{2},$$

then we have

$$\|\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\boldsymbol{w}^I{}_t(r-1))-g^a(\boldsymbol{w}^I{}_t(r-1))\|\leq rM_1^r\frac{1}{n}\mu.$$

Proof. First of all, $\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1))$ can be rewritten as below by using the Cauchy mean-value theorem:

$$\begin{aligned}\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1)) &= \frac{1}{n-r+1}[\sum_{i\notin R_{r-2}}\nabla F_i(\mathbf{w}^I{}_t(r-1))-\nabla F_{i_{r-1}}(\mathbf{w}^I{}_t(r-1))]\\ &= \frac{1}{n-r+1}\{(n-r+2)[\mathbf{H}_t^{r-2}\times(\mathbf{w}^I{}_t(r-1)-\mathbf{w}^I{}_t(r-2)]\\ &+\sum_{i\notin R_{r-2}}\nabla F_i(\mathbf{w}^I{}_t(r-2))-\nabla F_{i_{r-1}}(\mathbf{w}^I{}_t(r-1))\}.\end{aligned}$$

By subtracting the above formula from equation (62), i.e., the update rule for the approximate gradient, the norm of the approximation error between true and approximate gradients is:

$$\begin{aligned}&\|\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1))-g^a(\mathbf{w}^I{}_t(r-1))\|\\ &=\frac{1}{n-r+1}\|(n-r+2)(\mathbf{H}_t^{r-2}-\mathbf{B}_{j_m}^{r-2})\times(\mathbf{w}^I{}_t(r-1)-\mathbf{w}^I{}_t(r-2))\\ &+\sum_{i\notin R_{r-2}}\nabla F_i(\mathbf{w}^I{}_t(r-2))-(n-r+2)g^a(\mathbf{w}^I{}_t(r-2))\|\end{aligned}$$

Then by using the triangle inequality, Corollary 3 on the approximation accuracy of the quasi-Hessian (where the bound is in terms of ξ), and the bound on $\|\mathbf{w}^I{}_t(r-1)-\mathbf{w}^I{}_t(r-2)\|$, the formula above is bounded as:

$$\begin{aligned}&\leq\frac{n-r+2}{n-r+1}\|\mathbf{H}_t^{r-2}-\mathbf{B}_{j_m}^{r-2}\|\|\mathbf{w}^I{}_t(r-1)-\mathbf{w}^I{}_t(r-2)\|\\ &+\frac{1}{n-r+1}\|\sum_{i\notin R_{r-2}}\nabla F_i(\mathbf{w}^I{}_t(r-2))-(n-r+2)g^a(\mathbf{w}^I{}_t(r-2))\|\\ &\leq\frac{n-r+2}{n-r+1}\xi_{j_1,j_m}(r-2)M_1^r\frac{1}{n}+\frac{n-r+2}{n-r+1}\|\frac{1}{n-r+2}\sum_{i\notin R_{r-2}}\nabla F_i(\mathbf{w}^I{}_t(r-2))-g^a(\mathbf{w}^I{}_t(r-2))\|\end{aligned}\tag{68}$$

By using that $\xi_{j_1,j_m}(r-2) \leq \frac{\mu}{2}$, the formula above is bounded as:

$$\leq \frac{n-r+2}{n-r+1}\frac{\mu}{2}(M_1^r\frac{1}{n}) + \frac{n-r+2}{n-r+1}\|\frac{1}{n-r+2}\sum_{i\notin R_{r-2}}\nabla F_i(\mathbf{w}^I{}_t(r-2)) - g^a(\mathbf{w}^I{}_t(r-2))\|$$

We can use this recursively. Note that $\nabla F(\mathbf{w}^I{}_t(0)) = g^a(\mathbf{w}^I{}_t(0))$. In the end, we get the following inequality:

$$\leq \sum_{k=1}^{r-1}\frac{n-k}{n-r}\frac{\mu}{2}(M_1^r\frac{1}{n}) \leq M_1^r\frac{1}{n}\frac{\mu}{2}\sum_{k=1}^{r-1}\frac{n-k}{n-r}$$

Also for $r \ll n$, $\frac{n-k}{n-r} \leq 2$ (in fact we assumed $r/n \leq \delta$ for a sufficiently small δ, so this holds). So we get the bound $rM_1^r\frac{1}{n}\mu$.

□

Note that for $\|\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1)) - g^a(\mathbf{w}^I{}_t(r-1))\|$, we get a tighter bound when $t\to\infty$ by using equation (68), Lemma 11 (i.e. $d_{j_a,j_b}(r) \leq d_{j_a,j_b}(0) + 2r\cdot M_1^r\frac{1}{n}$) and Lemma 8 without using $\xi_{j_1,j_m}(r-1) \leq \frac{\mu}{2}$, which starts by bounding $\xi_{j_1,j_m}(k-1)$ where $k <= r$, $j_1 = j_0 + xT_0$ and $j_m = j_0 + (x+m-1)T_0$:

$$\begin{aligned}
&\xi_{j_1,j_m}(k-1) = Ad_{j_1,j_m+T_0-1}(k-1) + AM_1^r\frac{1}{n}\\
&\leq Ad_{j_1,j_m+T_0-1}(0) + 2(k-1)A\cdot M_1^r\frac{1}{n} + AM_1^r\frac{1}{n}\\
&\leq A(1-\mu\eta)^{j_0+xT_0}d_{0,mT_0-1}(0) + A(2k-1)M_1^r\frac{1}{n},
\end{aligned} \tag{69}$$

which can be plugged into Equation (68), i.e.:

$$\begin{aligned}
&\|\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1)) - g^a(\mathbf{w}^I{}_t(r-1))\|\\
&\leq \frac{n-r+2}{n-r+1}\xi_{j_1,j_m}(r-2)M_1^r\frac{1}{n}\\
&+ \frac{n-r+2}{n-r+1}\|\frac{1}{n-r+2}\sum_{i\notin R_{r-2}}\nabla F_i(\mathbf{w}^I{}_t(r-2)) - g^a(\mathbf{w}^I{}_t(r-2))\|\\
&\leq \sum_{k=1}^{r-1}\frac{n-k+1}{n-k}\xi_{j_1,j_m}(k-1)M_1^r\frac{1}{n}\\
&\leq \sum_{k=1}^{r-1}\frac{n-k+1}{n-k}[A(1-\mu\eta)^{j_0+xT_0}d_{0,mT_0-1}(0) + A(2k-1)M_1^r\frac{1}{n}]\cdot M_1^r\frac{1}{n}\\
&\leq 2A(1-\mu\eta)^{j_0+xT_0}d_{0,mT_0-1}(0)rM_1^r\frac{1}{n} + 2A(rM_1^r\frac{1}{n})^2
\end{aligned} \tag{70}$$

The last step uses that $\frac{n-k+1}{n-k} \leq 2$ and $\sum_{k=1}^{r-1}(2k-1) < \sum_{k=1}^{r}(2k-1) = r^2$. So when $t\to\infty$ and thus $x\to\infty$, $\|\frac{1}{n-r}\sum_{i\notin R_r}\nabla F_i(\mathbf{w}^I{}_t(r)) - g^a(\mathbf{w}^I{}_t(r))\| = o(\frac{r}{n})$.

Then based on Lemma 12 and 13, the bound on $\|g^a(\mathbf{w}^I{}_t(r))\|$ becomes:

$$\begin{aligned}
&\|g^a(\mathbf{w}^I{}_t(r-1))\|\\
&= \|g^a(\mathbf{w}^I{}_t(r-1)) - \frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1)) + \frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1))\|\\
&\leq \|g^a(\mathbf{w}^I{}_t(r-1)) - \frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1))\| + \|\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1))\|\\
&= rM_1^r\frac{1}{n}\mu + M_1^r\frac{1}{n}Lr + c_2 = (r\mu + Lr)M_1^r\frac{1}{n} + c_2
\end{aligned} \tag{71}$$

Main results

Theorem 14 (Bound between iterates on full data and incrementally updated ones (online deletions))**.** *Suppose that for any $k < r$, $\|\boldsymbol{w}^I{}_t(k) - \boldsymbol{w}^I{}_t(k-1)\| \leq M_1^r \frac{1}{n}$. At the r_{th} deletion request, consider an iteration t indexed with j_m for which $j_m \leq t < j_m + T_0 - 1$, and suppose that we are at the x-th iteration of full gradient updates, so $j_1 = j_0 + xT_0$, $j_m = j_0 + (m-1+x)T_0$. Suppose that we have the bounds $\|\boldsymbol{H}_t^{r-1} - \boldsymbol{B}_{j_m}^{r-1}\| \leq \xi_{j_1,j_m}(r-1) = Ad_{j_1,j_m+T_0-1}(r-1) + A(M_1^r \frac{1}{n})$ (where we recalled the definition of ξ) and*

$$\xi_{j_1,j_m}(r-1) = Ad_{j_1,j_m+T_0-1}(r-1) + A(M_1^r \frac{1}{n}) \leq \frac{\mu}{2}$$

for all iterations x. Then

$$\|\boldsymbol{w}^I{}_{t+1}(r) - \boldsymbol{w}^I{}_{t+1}(r-1)\| \leq M_1^r \frac{1}{n}.$$

Recall that c_0 is the Lipshitz constant of the Hessian, M_1 and A are defined in Theorem 12 and Corollary 3 respectively, and do not depend on t,

Then by using the same derivation as the proof of Theorem 4, we get the following results at the r_{th} deletion request.

Theorem 15 (Bound between iterates on full data and incrementally updated ones (all iterations, online deletion))**.** *At the deletion request r, if for all $k < r$, $\|\boldsymbol{w}^I{}_t(k) - \boldsymbol{w}^I{}_t(k-1)\| \leq M_1^r \frac{1}{n}$ holds, then for any $j_m < t < j_m + T_0 - 1$,*

$$\|\boldsymbol{w}^I{}_t(r) - \boldsymbol{w}^I{}_t(r-1)\| \leq M_1^r \frac{1}{n}$$

and

$$\|\boldsymbol{H}_t^{r-1} - \boldsymbol{B}_{j_m}^{r-1}\| \leq \xi_{j_1,j_m}(r-1) := Ad_{j_1,j_m+T_0-1}(r-1) + AM_1^r \frac{1}{n}$$

and

$$\|\frac{1}{n-r+1} \sum_{i \notin R_{r-1}} \nabla F_i(\boldsymbol{w}^I{}_t(r-1)) - g^a(\boldsymbol{w}^I{}_t(r-1))\| \leq rM_1^r \frac{1}{n}\mu$$

hold

Then by induction (the base case is similar to Theorem 4), we know that the following theorem holds for all iterations t:

Theorem 16 (Bound between iterates on full data and incrementally updated ones (all iterations, all deletion requests, online deletion))**.** *At the r_{th} deletion request, for any $j_m < t < j_m + T_0 - 1$,*

$$\|\boldsymbol{w}^I{}_t(r) - \boldsymbol{w}^I{}_t(r-1)\| \leq M_1^r \frac{1}{n}$$

and

$$\|\boldsymbol{H}_t^{r-1} - \boldsymbol{B}_{j_m}^{r-1}\| \leq \xi_{j_1,j_m}(r-1) := Ad_{j_1,j_m+T_0-1}(r-1) + AM_1^r \frac{1}{n}$$

and

$$\|\frac{1}{n-r+1} \sum_{i \notin R_{r-1}} \nabla F_i(\boldsymbol{w}^I{}_t(r-1)) - g^a(\boldsymbol{w}^I{}_t(r-1))\| \leq rM_1^r \frac{1}{n}\mu$$

hold

Then by induction (from the r_{th} deletion request to the 1_{st} deletion request), the following inequality holds:

$$\|\mathbf{w}^I{}_t(r) - \mathbf{w}^I{}_t(0)\| = \|\mathbf{w}^I{}_t(r) - \mathbf{w}_t\| \leq r \cdot M_1^r \frac{1}{n}$$

Then by using equation (64), the following inequality holds:

$$\begin{aligned}&\|\mathbf{w}^U{}_t(r) - \mathbf{w}^I{}_t(r-1)\| = \|\mathbf{w}^U{}_t(r) - \mathbf{w}_t + \mathbf{w}_t - \mathbf{w}^I{}_t(r-1)\| \\ &\leq \|\mathbf{w}^U{}_t(r) - \mathbf{w}_t\| + \|\mathbf{w}_t - \mathbf{w}^I{}_t(r-1)\| \\ &\leq M_1 \frac{r}{n} + (r-1)\cdot M_1^r \frac{1}{n} := M_2 \frac{r}{n}\end{aligned} \tag{72}$$

where M_2 is a constant which does not depend on t or k.
In the end, we get a similar result for the bound on $\|\mathbf{w}^I{}_t(r) - \mathbf{w}^U{}_t(r)\|$:

Theorem 17 (Convergence rate of DeltaGrad (online deletion))**.** *At the r_{th} deletion request, for all iterations t, the result $\boldsymbol{w}^I{}_t(r)$ of DeltaGrad, Algorithm 2, approximates the correct iteration values $\boldsymbol{w}^U{}_t(r)$ at the rate*

$$\|\boldsymbol{w}^U{}_t(r) - \boldsymbol{w}^I{}_t(r)\| = o(\frac{r}{n}).$$

So $\|\boldsymbol{w}^U{}_t(r) - \boldsymbol{w}^I{}_t(r)\|$ is of a lower order than $\frac{r}{n}$.

The proof of Theorem 14

Proof. Note that the approximated update rules for $\mathbf{w}^I{}_t$ at the r_{th} and the $(r-1)_{st}$ deletion request are:

$$\begin{aligned}&\mathbf{w}^I{}_{t+1}(r) = \mathbf{w}^I{}_t(r) - \frac{\eta}{n-r}\{(n-r+1)[\mathbf{B}_{j_m}^{r-1}(\mathbf{w}^I{}_t(r) - \mathbf{w}^I{}_t(r-1)) \\ &+ g^a(\mathbf{w}^I{}_t(r-1))] - \nabla F_{i_r}(\mathbf{w}^I{}_t(r))\}\end{aligned} \tag{73}$$

and

$$\begin{aligned}&\mathbf{w}^I{}_{t+1}(r-1) = \mathbf{w}^I{}_t(r-1) - \frac{\eta}{n-r+1}\{(n-r+2)[\mathbf{B}_{j_m}^{r-2}(\mathbf{w}^I{}_t(r-1) - \mathbf{w}^I{}_t(r-2)) \\ &+ g^a(\mathbf{w}^I{}_t(r-2))] - \nabla F_{i_{r-1}}(\mathbf{w}^I{}_t(r-1))\}.\end{aligned} \tag{74}$$

Note that since

$$\begin{aligned}g^a(\mathbf{w}^I{}_t(r-1)) = \frac{1}{n-r+1}\{&(n-r+2)[\mathbf{B}_{j_m}^{r-2}(\mathbf{w}^I{}_t(r-1) - \mathbf{w}^I{}_t(r-2)) \\ &+ g^a(\mathbf{w}^I{}_t(r-2))] - \nabla F_{i_{r-1}}(\mathbf{w}^I{}_t(r-1))\},\end{aligned}$$

then equation (74) can be rewritten as:

$$\begin{aligned}&\mathbf{w}^I{}_{t+1}(r-1) = \mathbf{w}^I{}_t(r-1) - \frac{\eta}{n-r+1}\{(n-r+2)[\mathbf{B}_{j_m}^{r-2}(\mathbf{w}^I{}_t(r-1) - \mathbf{w}^I{}_t(r-2)) \\ &+ g^a(\mathbf{w}^I{}_t(r-2))] - \nabla F_{i_{r-1}}(\mathbf{w}^I{}_t(r-1))\} \\ &= \mathbf{w}^I{}_t(r-1) - \eta g^a(\mathbf{w}^I{}_t(r-1)).\end{aligned} \tag{75}$$

Then by subtracting equation (74) from equation (75), the result becomes:

$$\begin{aligned}&\mathbf{w}^I{}_{t+1}(r) - \mathbf{w}^I{}_{t+1}(r-1) \\ &= (\mathbf{w}^I{}_t(r) - \mathbf{w}^I{}_t(r-1)) - \frac{\eta}{n-r}\{(n-r+1)[\mathbf{B}_{j_m}^{r-1}(\mathbf{w}^I{}_t(r) - \mathbf{w}^I{}_t(r-1)) \\ &+ g^a(\mathbf{w}^I{}_t(r-1))] - \nabla F_{i_r}(\mathbf{w}^I{}_t(r))\} + \eta g^a(\mathbf{w}^I{}_t(r-1)).\end{aligned}$$

Then by adding and subtracting $\mathbf{H}_t^{r-1}$ and $\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F(\mathbf{w}^I{}_t(r-1))$ in the formula above and rearranging the result properly, it becomes:

$$\begin{aligned}
&\mathbf{w}^I{}_{t+1}(r)-\mathbf{w}^I{}_{t+1}(r-1)\\
&=(\mathbf{I}-\eta\frac{n-r+1}{n-r}(\mathbf{B}_{j_m}^{r-1}-\mathbf{H}_t^{r-1}))(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1))\\
&-\frac{\eta}{n-r}\{(n-r+1)[\mathbf{H}_t^{r-1}(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1))\\
&+\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F(\mathbf{w}^I{}_t(r-1))-\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F(\mathbf{w}^I{}_t(r-1))\\
&+g^a(\mathbf{w}^I{}_t(r-1))]-\nabla F_{i_r}(\mathbf{w}^I{}_t(r))\}+\eta g^a(\mathbf{w}^I{}_t(r-1)).
\end{aligned}\tag{76}$$

Then by using the fact that

$$\begin{aligned}
&\mathbf{H}_t^{r-1}(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1))+\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F(\mathbf{w}^I{}_t(r-1))\\
&=\frac{1}{n-r+1}\sum_{i\notin R_{r-1}}\nabla F(\mathbf{w}^I{}_t(r))
\end{aligned}$$

and

$$(\sum_{i\notin R_{r-1}}\nabla F(\mathbf{w}^I{}_t(r)))-\nabla F_{i_r}(\mathbf{w}^I{}_t(r))=\sum_{i\notin R_r}\nabla F(\mathbf{w}^I{}_t(r)),$$

Equation (76) becomes:

$$\begin{aligned}
&\mathbf{w}^I{}_{t+1}(r)-\mathbf{w}^I{}_{t+1}(r-1)\\
&=(\mathbf{I}-\eta\frac{n-r+1}{n-r}(\mathbf{B}_{j_m}^{r-1}-\mathbf{H}_t^{r-1}))(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1))\\
&-\frac{\eta}{n-r}[\sum_{i\notin R_r}\nabla F(\mathbf{w}^I{}_t(r))-\sum_{i\notin R_{r-1}}\nabla F(\mathbf{w}^I{}_t(r-1))\\
&+(n-r+1)g^a(\mathbf{w}^I{}_t(r-1))]+\eta g^a(\mathbf{w}^I{}_t(r-1)).
\end{aligned}\tag{77}$$

Also note that by using the Cauchy mean-value theorem, the following equation holds:

$$\begin{aligned}
&\sum_{i\notin R_r}\nabla F_i(\mathbf{w}^I{}_t(r))-\sum_{i\notin R_{r-1}}\nabla F_i(\mathbf{w}^I{}_t(r-1))\\
&=\sum_{i\notin R_r}\nabla F_i(\mathbf{w}^I{}_t(r))-\sum_{i\notin R_r}\nabla F_i(\mathbf{w}^I{}_t(r-1))-\nabla F_{i_r}(\mathbf{w}^I{}_t(r-1))\\
&=[\sum_{i\notin R_r}\int_0^1\mathbf{H}_i(\mathbf{w}^I{}_t(r-1)+x(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1)))dx](\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1))-\nabla F_{i_r}(\mathbf{w}^I{}_t(r-1)),
\end{aligned}$$

which can be plugged into equation (77), i.e.:

$$\begin{aligned}
&\mathbf{w}^I{}_{t+1}(r)-\mathbf{w}^I{}_{t+1}(r-1)\\
&=(\mathbf{I}-\eta\frac{n-r+1}{n-r}(\mathbf{B}_{j_m}^{r-1}-\mathbf{H}_t^{r-1}))(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1))\\
&-\frac{\eta}{n-r}\{[\sum_{i\notin R_r}\int_0^1\mathbf{H}_i(\mathbf{w}^I{}_t(r-1)+x(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1)))dx]\\
&\cdot(\mathbf{w}^I{}_t(r)-\mathbf{w}^I{}_t(r-1))-\nabla F_{i_r}(\mathbf{w}^I{}_t(r-1))\\
&+(n-r+1)g^a(\mathbf{w}^I{}_t(r-1))\}+\eta g^a(\mathbf{w}^I{}_t(r-1)),
\end{aligned}\tag{78}$$

which can be rearranged as:

$$\begin{aligned}
&\mathbf{w^I}_{t+1}(r) - \mathbf{w^I}_{t+1}(r-1) \\
&= (\mathbf{I} - \eta\frac{n-r+1}{n-r}(\mathbf{B}_{j_m}^{r-1} - \mathbf{H}_t^{r-1}))(\mathbf{w^I}_t(r) - \mathbf{w^I}_t(r-1)) \\
&- \frac{\eta}{n-r}\{[\sum_{i\notin R_r}\int_0^1 \mathbf{H}_i(\mathbf{w^I}_t(r-1) + x(\mathbf{w^I}_t(r) - \mathbf{w^I}_t(r-1)))dx] \\
&\cdot (\mathbf{w^I}_t(r) - \mathbf{w^I}_t(r-1)) - \nabla F_{i_r}(\mathbf{w^I}_t(r-1))\} - \frac{\eta}{n-r}g^a(\mathbf{w^I}_t(r-1)).
\end{aligned} \tag{79}$$

Then by taking the matrix norm on both sides of equation (79) and using that $\|\mathbf{H}_i(\mathbf{w^I}_t(r-1)+x(\mathbf{w^I}_t(r)-\mathbf{w^I}_t(r-1)))\|\geq \mu$ and $\|\mathbf{B}_{j_m}^{r-1} - \mathbf{H}_t^{r-1}\|\leq \xi_{j_1,j_m}(r-1)$, equation (79) can be bounded as:

$$\begin{aligned}
&\|\mathbf{w^I}_{t+1}(r) - \mathbf{w^I}_{t+1}(r-1)\| \\
&\leq (1-\eta\mu)\|\mathbf{w^I}_t(r) - \mathbf{w^I}_t(r-1)\| \\
&+ \frac{(n-r+1)\eta}{n-r}\xi_{j_1,j_m}(r-1)\|\mathbf{w^I}_t(r) - \mathbf{w^I}_t(r-1)\| \\
&+ \frac{\eta}{n-r}\|\nabla F_{i_r}(\mathbf{w^I}_t(r-1))\| + \|\frac{\eta}{n-r}g^a(\mathbf{w^I}_t(r-1))\|.
\end{aligned}$$

Then by using Lemma 12 and equation (71), the formula above becomes:

$$\begin{aligned}
&\leq (1 - \eta\mu + \frac{(n-r+1)\eta}{n-r}\xi_{j_1,j_m}(r))\|\mathbf{w^I}_t(r) - \mathbf{w^I}_t(r-1)\| \\
&+ \frac{\eta}{n-r}(M_1^r\frac{1}{n}L(r-1) + c_2) + \frac{\eta}{n-r}(M_1^r\frac{1}{n}(r-1)\mu + M_1^r\frac{1}{n}L(r-1) + c_2).
\end{aligned}$$

By using the bound on $\xi_{j_1,j_m}(r)$ and applying the above formula recursively across all iterations, the formula above becomes:

$$\begin{aligned}
&\leq \frac{1}{\eta\mu - \frac{\eta(n-r+1)}{n-r}\frac{\mu}{2}}(\frac{\eta}{n-r}(M_1^r\frac{1}{n}L(r-1) + c_2) \\
&+ \frac{\eta}{n-r}(M_1^r\frac{1}{n}L(r-1) + M_1^r\frac{1}{n}\mu(r-1) + c_2) \\
&= \frac{2}{(n-r-1)\mu}((M_1^r\frac{1}{n}(2L(r-1) + (r-1)\mu) + 2c_2).
\end{aligned}$$

Then by using that $M_1 = \frac{2c_2}{\mu}$ and $M_1^r\frac{1}{n} = \frac{\frac{2M_1}{n}}{1-\frac{r+1}{n} - \frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})}$, the formula above can be rewritten as:

$$\begin{aligned}
&= \frac{2}{(n-r-1)\mu}\frac{\frac{2M_1(r-1)}{n}(2L+\mu) + \mu M_1(1 - \frac{r+1}{n} - \frac{2(r-1)}{n}(\frac{2L+\mu}{\mu}))}{1 - \frac{r+1}{n} - \frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})} \\
&= \frac{2\frac{M_1}{n}}{1 - \frac{r+1}{n} - \frac{2(r-1)}{n}(\frac{2L+\mu}{\mu})} = M_1^r\frac{1}{n}.
\end{aligned}$$

This finishes the proof.

□

The proof of Theorem 17

Proof. Recall that the update rule for $\mathbf{w^U}_t(r)$ is:

$$\mathbf{w^U}_{t+1}(r) = \mathbf{w^U}_t(r) - \eta\frac{1}{n-r}\sum_{i\notin R_r}\nabla F(\mathbf{w^U}_t(r))$$

and the update rule for ${\mathbf{w}^I}_t(r)$ is (where the gradients are explicitly evaluated):

$${\mathbf{w}^I}_{t+1}(r) = {\mathbf{w}^I}_t(r) - \frac{\eta}{n-r}[(n-r+1)(\mathbf{B}_{j_m}^{r-1}({\mathbf{w}^I}_t(r) - {\mathbf{w}^I}_t(r-1)) + g^a({\mathbf{w}^I}_t(r-1))) - \nabla F_{i_r}({\mathbf{w}^I}_t(r))].$$

Then by subtracting ${\mathbf{w}^I}_{t+1}(r)$ from ${\mathbf{w}^U}_{t+1}(r)$, we get:

$$\begin{aligned}
&\|{\mathbf{w}^I}_{t+1}(r) - {\mathbf{w}^U}_{t+1}(r)\| \\
&= \|{\mathbf{w}^I}_t(r) - {\mathbf{w}^U}_t(r) - \frac{\eta}{n-r}\{(n-r+1)[\mathbf{B}_{j_m}^{r-1}({\mathbf{w}^I}_t(r) - {\mathbf{w}^I}_t(r-1)) \\
&+ g^a({\mathbf{w}^I}_t(r-1))] - \nabla F_{i_r}({\mathbf{w}^I}_t(r))\} + \frac{\eta}{n-r}\sum_{i \notin R_r} \nabla F({\mathbf{w}^U}_t(r))\|.
\end{aligned}$$

Then by bringing in $\mathbf{H}_t^{r-1}$ and $\frac{1}{n-r+1}\sum_{i \in R_{r-1}} \nabla F({\mathbf{w}^I}_t(r-1))$ into the formula above, we get:

$$\begin{aligned}
&= \|{\mathbf{w}^I}_t(r) - {\mathbf{w}^U}_t(r) - \frac{(n-r+1)\eta}{n-r}\left[\left(\mathbf{B}_{j_m}^{r-1} - \mathbf{H}_t^{r-1}\right)({\mathbf{w}^I}_t(r) - {\mathbf{w}^I}_t(r-1))\right. \\
&+ \mathbf{H}_t^{r-1} \times ({\mathbf{w}^I}_t(r) - {\mathbf{w}^I}_t(r-1)) + g^a({\mathbf{w}^I}_t(r-1)) \\
&- \frac{1}{n-r+1}\sum_{i \in R_{r-1}} \nabla F({\mathbf{w}^I}_t(r-1)) + \frac{1}{n-r+1}\sum_{i \in R_{r-1}} \nabla F({\mathbf{w}^I}_t(r-1))] \\
&+ \frac{\eta}{n-r}[\nabla F_{i_r}({\mathbf{w}^I}_t(r)) - \nabla F_{i_r}({\mathbf{w}^I}_t(r-1)) + \nabla F_{i_r}({\mathbf{w}^I}_t(r-1))] + \frac{\eta}{n-r}\sum_{i \notin R_r} \nabla F_i({\mathbf{w}^U}_t(r))\|.
\end{aligned}$$

Then by using the triangle inequality and the result from equation (70), the formula above can be bounded as:

$$\begin{aligned}
&\leq \|{\mathbf{w}^I}_t(r) - {\mathbf{w}^U}_t(r) - \frac{(n-r+1)\eta}{n-r}\left[\left(\mathbf{B}_{j_m}^{r-1} - \mathbf{H}_t^{r-1}\right)({\mathbf{w}^I}_t(r) - {\mathbf{w}^I}_t(r-1))\right. \\
&\left. + \mathbf{H}_t^{r-1} \times ({\mathbf{w}^I}_t(r) - {\mathbf{w}^I}_t(r-1)) + \frac{1}{n-r+1}\sum_{i \in R_{r-1}} \nabla F_i({\mathbf{w}^I}_t(r-1))\right] \\
&+ \frac{\eta}{n-r}[\nabla F_{i_r}({\mathbf{w}^I}_t(r)) - \nabla F_{i_r}({\mathbf{w}^I}_t(r-1)) + \nabla F_{i_r}({\mathbf{w}^I}_t(r-1))] \\
&+ \frac{\eta}{n-r}\sum_{i \notin R_r} \nabla F_i({\mathbf{w}^U}_t(r))\| + 2A(1-\mu\eta)^{j_0+xT_0} d_{0,(m-1)T_0}(0) r M_1^r \frac{1}{n} + 2A(rM_1^r\frac{1}{n})^2.
\end{aligned}$$

Note that the first matrix norm in this formula is the same as equation (33) by replacing n, r, ${\mathbf{w}^I}_t$, ${\mathbf{w}^U}_t$, $\mathbf{w}_t$, $\mathbf{B}_{j_m}$, $\mathbf{H}_t$ and $\nabla F(\mathbf{w}_t)$ with $n-r+1$, 1, ${\mathbf{w}^I}_t(r)$, ${\mathbf{w}^U}_t(r)$, ${\mathbf{w}^I}_t(r-1)$, $\mathbf{B}_{j_m}^{r-1}$, $\mathbf{H}_t^{r-1}$ and $\frac{1}{n-r+1}\sum_{i \notin R_{r-1}} \nabla F_i({\mathbf{w}^I}_t(r-1))$ reps.. So by following the same derivation, the formula above can be bounded as:

$$\begin{aligned}
&\leq \|(\mathbf{I} - \frac{\eta}{n-r}\sum_{i \notin R_r} \mathbf{H}_{t,i}^{r-1})({\mathbf{w}^I}_t(r) - {\mathbf{w}^U}_t(r))\| \\
&+ \|\frac{(n-r+1)\eta}{n-r}\left[\left(\mathbf{B}_{j_m}^{r-1} - \mathbf{H}_t^{r-1}\right)({\mathbf{w}^I}_t(r) - {\mathbf{w}^U}_t(r))\right]\| \\
&+ \|\frac{\eta}{n-r}[\sum_{i \notin R_r}\int_0^1 \mathbf{H}_i({\mathbf{w}^I}_t(r-1) + x({\mathbf{w}^U}_t(r) - {\mathbf{w}^I}_t(r-1)))dx \\
&- \int_0^1 \mathbf{H}_i({\mathbf{w}^I}_t(r-1) + x({\mathbf{w}^I}_t(r) - {\mathbf{w}^I}_t(r-1)))dx]({\mathbf{w}^U}_t(r) - {\mathbf{w}^I}_t(r-1))\| \\
&+ \|\frac{(n-r+1)\eta}{n-r}\left[\left(\mathbf{B}_{j_m}^{r-1} - \mathbf{H}_t^{r-1}\right)({\mathbf{w}^U}_t(r) - {\mathbf{w}^I}_t(r-1))\right]\| \\
&+ 2A(1-\mu\eta)^{j_0+xT_0} d_{0,(m-1)T_0}(0) r M_1^r \frac{1}{n} + 2A(rM_1^r\frac{1}{n})^2.
\end{aligned}$$

Then by using the following facts:

1. $\|\mathbf{I} - \eta \mathbf{H}_{t,i}^{r-1}\| \leq 1 - \eta\mu$;
2. from Theorem 16 on the approximation accuracy of the quasi-Hessian to mean Hessian, we have the error bound $\|\mathbf{H}_t^{r-1} - \mathbf{B}_{j_m}^{r-1}\| \leq \xi_{j_1,j_m}(r-1)$;
3. we bound the difference of integrated Hessians using the strategy from Equation (20);
4. from Equation (72), we have the error bound $\|\mathbf{w}^U{}_t(r) - \mathbf{w}^I{}_t(r-1)\| \leq M_2 \frac{r}{n}$ (and this requires no additional assumptions),

the expression can be bounded as follows:

$$
\begin{aligned}
&\leq (1 - \eta\mu + \frac{(n-r+1)\eta}{n-r}\xi_{j_0,j_0+(m-1)T_0}(r-1) + \frac{c_0 M_2 r\eta}{2n})\|\mathbf{w}^I{}_t - \mathbf{w}^U{}_t\| \\
&+ \frac{M_2(n-r+1)r\eta}{n(n-r)}\xi_{j_1,j_m}(r-1) + 2A(1-\mu\eta)^{j_0+xT_0} d_{0,(m-1)T_0}(0) r M_1^r \frac{1}{n} \\
&+ 2A(rM_1^r \frac{1}{n})^2,
\end{aligned}
$$

which is very similar to equation (36) (except the difference in the coefficient). So by following the derivation after equation (36), we know that:

$$\|\mathbf{w}^I{}_t(r) - \mathbf{w}^U{}_t(r)\| = o(\frac{r}{n})$$

when $t \to \infty$.

□

C.3 Extension of DeltaGrad for non-strongly convex, non-smooth objective functions

For the original version of the L-BFGS algorithm, strong convexity is essential to make the secant condition hold. In this subsection, we present our extension of DeltaGrad to non-strongly convex, non-smooth objectives.

To deal with non-strongly convex objectives, we assume that convexity holds in some local regions. When constructing the arrays ΔG and ΔW, only the model parameters and their gradients where local convexity holds are used.

For local non-smoothness, we found that even a small distance between $\mathbf{w}_t$ and $\mathbf{w}^I{}_t$ can make the estimated gradient $\nabla F(\mathbf{w}^I{}_t)$ drift far away from $\nabla F(\mathbf{w}_t)$. To deal with this, we explicitly check if the norm of $\mathbf{B}_{j_m}(\mathbf{w}_t - \mathbf{w}^I{}_t)$ (which equals to $\nabla F(\mathbf{w}^I{}_t) - \nabla F(\mathbf{w}_t)$) is larger than the norm of $L(\mathbf{w}_t - \mathbf{w}^I{}_t)$ for a constant L. In our experiments, L is configured as 1. The details of the modifications above are highlighted in Algorithm 3.

D Supplementary experiments

In this section, we present some supplementary experiments that could not be presented in the paper due to space limitations.

D.1 Experiments with large deletion rate

In this experiment, instead of deleting at most 1% of training samples each time as we did in Section 4 in the main paper, we vary the deletion rate from 0 to up to 20% on MNIST dataset and still compare the performance between DeltaGrad (with T_0 as 5 and j_0 as 10) and BaseL. All other hyper-parameters such as the learning rate and mini-batch size remain the same as in Section 4 in the main paper.

The experimental results in Figure 1 show that even with the largest deletion rate, i.e. 20%, DeltaGrad can still be 1.67x faster than BaseL (2.27s VS 1.53s) and the error bound between their resulting model

Algorithm 3: DeltaGrad (general models)

Input : The full training set $(\mathbf{X}, \mathbf{Y})$, model parameters cached during the training phase for the full training samples $\{\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t\}$ and corresponding gradients $\{\nabla F(\mathbf{w}_0), \nabla F(\mathbf{w}_1), \ldots, \nabla F(\mathbf{w}_t)\}$, the removed training sample or the added training sample R, period T_0, total iteration number T, history size m, warmup iteration number j_0, learning rate η

Output: Updated model parameter ${\mathbf{w}^I}_t$

1 Initialize ${\mathbf{w}^I}_0 \leftarrow \mathbf{w}_0$
2 Initialize an array $\Delta G = []$
3 Initialize an array $\Delta W = []$
4 Initialize $last\_t = j_0$
5 $is\_explicit = False$
6 **for** $t = 0; t < T; t++$ **do**
7 **if** $(t - last_t) \mod T_0 == 0$ *or* $t \leq j_0$ **then**
8 $is\_explicit = True$
9 **else**
10 **end**
11 **if** $is\_explicit == True$ *or* $t \leq j_0$ **then**
12 $last\_t = t$
13 compute $\nabla F\left({\mathbf{w}^I}_t\right)$ exactly
14 compute $\nabla F\left({\mathbf{w}^I}_t\right) - \nabla F(\mathbf{w}_t)$ based on the cached gradient $\nabla F(\mathbf{w}_t)$
 `/* check local convexity */`
15 **if** $< \nabla F\left({\boldsymbol{w}^I}_t\right) - \nabla F(\boldsymbol{w}_t), {\boldsymbol{w}^I}_t - \boldsymbol{w}_t > \leq 0$ **then**
16 compute ${\mathbf{w}^I}_{t+1}$ by using exact GD update (equation (1))
17 **continue**
18 **end**
19 set $\Delta G[k] = \nabla F\left({\mathbf{w}^I}_t\right) - \nabla F(\mathbf{w}_t)$
20 set $\Delta W[k] = {\mathbf{w}^I}_t - \mathbf{w}_t$, based on the cached parameters $\mathbf{w}_t$
21 $k \leftarrow k + 1$
22 compute ${\mathbf{w}^I}_{t+1}$ by using exact GD update (equation (1))
23 **else**
24 Pass $\Delta W[-m:]$, $\Delta G[-m:]$, the last m elements in ΔW and ΔG, which are from the $j_1^{th}, j_2^{th}, \ldots, j_m^{th}$ iterations where $j_1 < j_2 < \cdots < j_m$ depend on t, $\mathbf{v} = {\mathbf{w}^I}_t - \mathbf{w}_t$, and the history size m, to the L-BFGFS Algorithm (See Supplement) to get the approximation of $\mathbf{H}(\mathbf{w}_t)\mathbf{v}$, i.e., $\mathbf{B}_{j_m}\mathbf{v}$
 `/* check local smoothness */`
25 **if** $\|\boldsymbol{B}_{j_m}\boldsymbol{v}\| \geq \|\boldsymbol{v}\|$ **then**
26 **go to** line 12
27 **end**
28 Approximate $\nabla F\left({\mathbf{w}^I}_t\right) = \nabla F(\mathbf{w}_t) + \mathbf{B}_{j_m}\left({\mathbf{w}^I}_t - \mathbf{w}_t\right)$
29 Compute ${\mathbf{w}^I}_{t+1}$ by using the "leave-r-out" gradient formula, based on the approximated $\nabla F({\mathbf{w}^I}_t)$
30 **end**
31 **end**
32 **return** ${\mathbf{w}^I}_t$

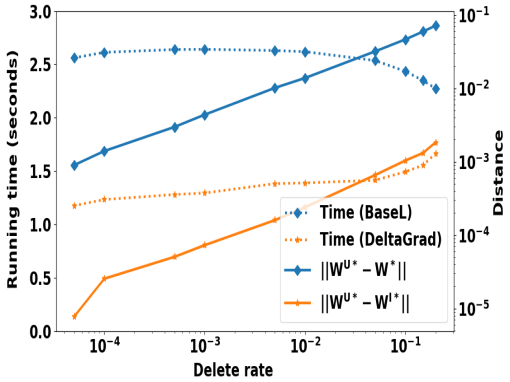


Figure 1: Running time and distance with varied deletion rate up to 20%

parameters (i.e. $\mathbf{w}^{I*}$ VS $\mathbf{w}^{U*}$) are still acceptable (on the order of 10^{-3}), far smaller than the error bound between $\mathbf{w}^{U*}$ and $\mathbf{w}^{*}$ (on the order of 10^{-1}). Such a small difference between $\mathbf{w}^{I*}$ and $\mathbf{w}^{U*}$ also results in almost the same prediction performance, i.e. $87.460 \pm 0.0011\%$ and $87.458 \pm 0.0012\%$ respectively. This experiment thus provides some justification for the feasibility of DeltaGrad even when the number of the removed samples is not far smaller than the entire training dataset size.

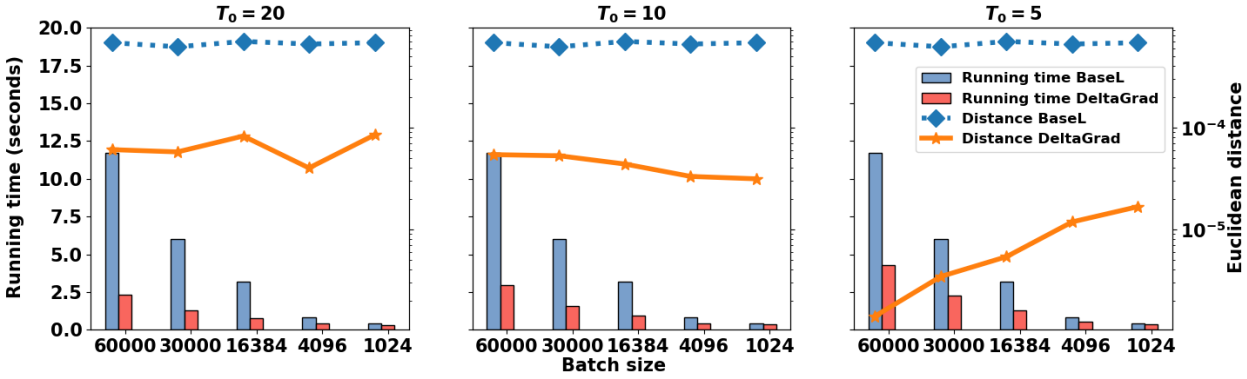


Figure 2: Running time and distance comparison with varying mini-batch size under fixed $j_0 = 10$ and varying T_0 ($T_0 = 20$ VS $T_0 = 10$ VS $T_0 = 5$)

D.2 Influence of hyper-parameters on performance

To begin with, the influence of different hyper-parameters used in SGD and DeltaGrad is explored. We delete one sample from the training set of MNIST by running regularized logistic regression with the same learning rate and regularization rate as in Section 4 and varying mini-batch sizes (1024 - 60000), T_0 ($T_0 = 20$, 10, 5) and j_0 ($j_0 = 5$, 10, 50). The experimental results are presented in Figure 2-3. For different mini-batch sizes, we also used different epoch numbers to make sure that the total number of running iterations/steps in SGD are roughly the same. In what follows, we analyze how the mini-batch size, the hyper-parameters T_0 and j_0 influence the performance, thus providing some hints on how to choose proper hyper-parameters when DeltaGrad is used.

Influence of the mini-batch size. It is clear from Figure 2-3 that with larger mini-batch sizes, DeltaGrad can gain more speed with longer running time for both BaseL and DeltaGrad. As discussed in Section 4, to compute the gradients, other GPU-related overhead (the overhead to copy data from CPU

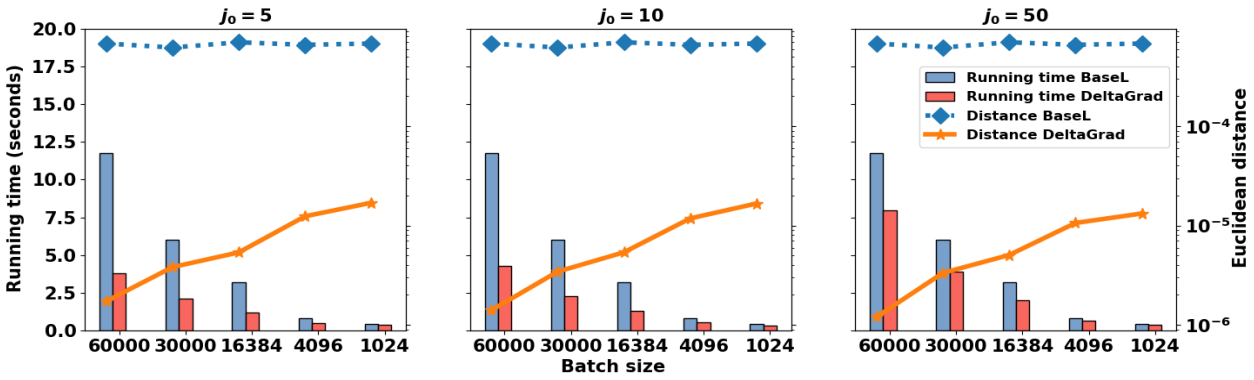


Figure 3: Running time and distance comparison with varying mini-batch size under fixed $T_0 = 5$ and varying j_0 ($j_0 = 5$ VS $j_0 = 10$ VS $j_0 = 50$)

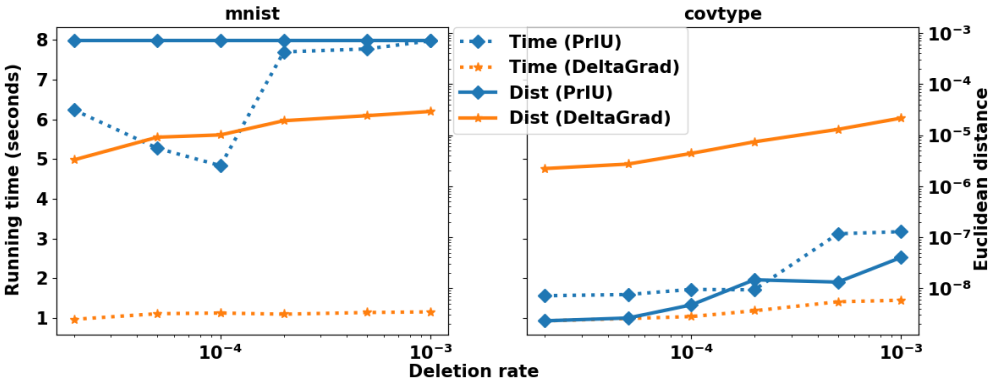


Figure 4: Comparison of DeltaGrad and PrIU

DRAM to GPU DRAM, the time to launch the kernel on GPU) cannot be ignored. This can become more significant when compared against the smaller computational overhead for smaller mini-batch data. Also notice that, when $T_0 = 5$, with increasing B, the difference between $\mathbf{w}^U$ and $\mathbf{w}^I$ becomes smaller and smaller, which matches our conclusion in Theorem 11, i.e. with larger B, the difference $o(\frac{r}{n} + \frac{1}{B^{\frac{1}{4}}})$ is smaller.

Influence of T_0. By comparing the three sub-figures in Figure 2, the running time slightly (rather than significantly) decreases with increasing T_0 for the same mini-batch size. This is explained by the earlier analysis in Section 4 on the non-ideal performance for GPU computation over small matrices. Interestingly, when $T_0 = 10$ or $T_0 = 20$, $\|\mathbf{w}^{I,S} - \mathbf{w}^{U,S}\|$ does not decrease with larger mini-batch sizes. This is because in Formula (61), one component of the bound of $\|\mathbf{w}^{I,S} - \mathbf{w}^{U,S}\|$ is

$$\frac{M_1(\frac{r}{n} + \frac{1}{B^{1/4}})}{C(1 - \frac{r}{n} - \frac{1}{B^{1/4}})}(1 - \eta C)^{yT_0}(1 - \eta\mu)^{j_0} d_{0,mT_0-1} \frac{1}{1 - (\frac{1-\eta\mu}{1-\eta C})^{T_0}}$$

(while the other component is $o((\frac{r}{n} + \frac{1}{B^{\frac{1}{4}}}))$). Here d_{0,mT_0-1} increases with larger T_0 and the term $(1-\eta C)^{yT_0}$ is not arbitrarily approaching 0 since yT_0 cannot truly go to infinity. So when $T_0 = 20$ and $T_0 = 10$, this component becomes the dominating term in the bound of $\|\mathbf{w}^{I,S} - \mathbf{w}^{U,S}\|$. So to make the bound $o((\frac{r}{n} + \frac{1}{B^{\frac{1}{4}}}))$ hold, so that we can adjust the bound of $\|\mathbf{w}^{I,S} - \mathbf{w}^{U,S}\|$ by varying B, proper choice of T_0 is important. For example, $T_0 = 5$ is a good choice for the MNIST dataset. This can achieve speed-ups comparable to larger T_0 without sacrificing the closeness between $\mathbf{w}^{I,S}$ and $\mathbf{w}^{U,S}$.

Influence of j_0. By comparing the three sub-figures in Figure 3, with increasing j_0, long "burn-in"

iterations are expected, thus incurring more running time. This, however, does not significantly reduce the distance between $\mathbf{w}^{I,S}$ and $\mathbf{w}^{U,S}$. It indicates that we can select smaller j_0, e.g. 5 or 10 for more speed-up.

Discussions on tuning the hyper-parameters for DeltaGrad. Through our extensive experiments, we found that for regularized logistic regression, setting T_0 as 5 and j_0 as $5-20$ would lead to some of the most favorable trade-offs between running time and the error $\|\mathbf{w}^{U,S} - \mathbf{w}^{I,S}\|$. But in terms of more complicated models, e.g. 2-layer DNN, higher j_0 (even half of the total iteration number) and smaller T_0 (2 or 1) are necessary. Similar experiments were also conducted on adding training samples, in which similar trends were observed.

D.3 Comparison against the state-of-the-art work

To our knowledge, the closest work to ours is Wu et al. (2020), which targets simple ML models, i.e. linear regression and regularized logistic regression with an ad-hoc solution (called PrIU) rather than solutions for general models. Their solutions can only deal with the deletion of samples from the training set without supporting the addition of samples. In our experiments, we compared DeltaGrad (with $T_0 = 5$ and $j_0 = 10$) against PrIU by running regularized logistic regression over MNIST and covtype with the same mini-batch size (16384), the same learning rate and regularization rate, but with varying deletion rates.

Table 1: Memory usage of DeltaGrad and PrIU(GB)

Deletion rate	MNIST		covtype	
	PrIU	DeltaGrad	PrIU	DeltaGrad
2×10^{-5}	26.61	2.74	9.30	2.56
5×10^{-5}	27.02	2.74	9.30	2.56
1×10^{-4}	27.13	2.74	9.30	2.55
2×10^{-4}	27.75	2.74	9.31	2.56
5×10^{-4}	29.10	2.74	10.67	2.56
1×10^{-3}	29.10	2.74	10.67	2.56

The running time and the distance term $\|\mathbf{w}^U - \mathbf{w}^I\|$ of both PrIU and DeltaGrad with varying deletion rate are presented in Figure 4. First, it shows that DeltaGrad is always faster than PrIU, with more significant speed-ups on MNIST. The reason is that the time complexity of PrIU is $O(rp)$ for each iteration where p represents the total number of model parameters while r represents the reduced dimension after Singular Value Decomposition is conducted over some $p \times p$ matrix. This is a large integer for large sparse matrices, e.g. MNIST.

As a result, $O(rp)$ is larger than the time complexity of DeltaGrad. Also, the memory usage of PrIU and DeltaGrad is shown in Table 1. PrIU needs much more DRAM (even 10x in MNIST) than DeltaGrad. The reason is that to prepare for the model update phase, PrIU needs to collect more information during the training phase over the full dataset. This is needed in the model update phase and is quadratic in the number of the model parameters p. The authors of Wu et al. (2020) claimed that their solution cannot provide good performance over sparse datasets in terms of running time, error term $\mathbf{w}^U - \mathbf{w}^I$ and memory usage. In contrast, both the time and space overhead of DeltaGrad are smaller, which thus indicates the potential of its usage in the realistic, large-scale scenarios.

D.4 Experiments on large ML models

In this section, we compare DeltaGrad with BaseL using the state-of-the-art ResNet152 network (He et al., 2016) (ResNet for short hereafter) with all but the top layer frozen, for which we use the pre-trained parameters from Pytorch torchvision library[1]. The pre-trained layers with fixed parameters are regarded as the feature transformation layer, applied over each training sample as the pre-processing step before the training phase. Those transformed features are then used to train the last layer of ResNet, which is thus equivalent to training a logistic regression model.

This experiment is conducted on CIFAR-10 dataset (Krizhevsky et al., 2009), which is composed of 60000 32×32 color images (50000 of them are training samples while the rest of are test samples). We run

[1]https://pytorch.org/docs/stable/torchvision/models.html

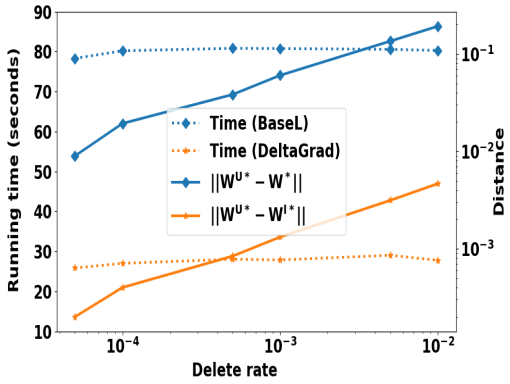


Figure 5: Comparison of DeltaGrad and BaseL on the CIFAR-10 dataset with pre-trained ResNet152 network

SGD with mini-batch size 10000, fixed learning rate 0.05 and L2 regularization rate 0.0001. Similar to the experimental setup introduced in Section 4 in the main paper, the deletion rate is varied from 0 to 1% and the model parameters are updated by using BaseL and DeltaGrad (with T_0 as 5 and j_0 as 20) respectively after the deletion operations. The experimental results are presented in Figure 5, again showing significant speed-ups for DeltaGrad relative to BaseL (up to 3x speed-ups when the deletion rate is 0.005%) with far smaller error bound (up to 4×10^{-3}) than the baseline error bound (up to 2×10^{-2}). Since it is quite common to reuse sophisticated pre-trained models in practice, we expect that the use of DeltaGrad in this manner is applicable in many cases.

D.5 Applications of DeltaGrad to robust learning

As Section 5 in the main paper reveals, DeltaGrad has many potential applications. In this section, we explored how DeltaGrad can accelerate the evaluations of the effect of the outliers in robust statistical learning. Here the effect of outliers is represented by the difference of the model parameters before and after the deletion of the outliers (see Yu and Yao (2017)).

In the experiments, we start by training a model on the training dataset (RCV1 here) along with some randomly generated outliers. Then we remove those outliers and update the model on the remaining training samples by using DeltaGrad and BaseL. We also evaluate the effect of the fraction of outlier: the ratio between the number of the outliers and the training dataset size is also defined as the *Deletion rate*. It is varied from 1% to 10%. According to the experimental results shown in Figure 6, when there are up to 10% outliers in the training dataset, DeltaGrad is at least 2.18x faster than BaseL in evaluating the updated model parameters by only sacrificing little computational accuracy (no more than 5×10^{-3}), thus reducing the computational overhead on evaluating the effect of the outliers in robust learning.

References

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60 (2):223–311, 2018.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3):129–156, 1994.

K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *Advances in neural information processing systems*, pages 289–296, 2009.

A. R. Conn, N. I. Gould, and P. L. Toint. Convergence of quasi-newton matrices generated by the symmetric rank one update. *Mathematical programming*, 50(1-3):177–195, 1991.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making ai forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, pages 3513–3526, 2019.

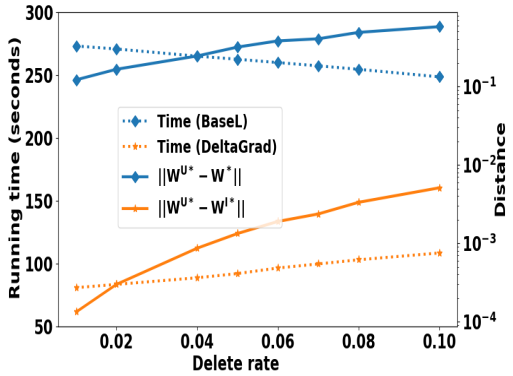


Figure 6: Comparison of DeltaGrad and BaseL on RCV1 dataset after deleting outliers

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

R. I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

J. A. Tropp. The expected norm of a sum of independent random matrices: An elementary approach. In *High dimensional probability VII*, pages 173–202. Springer, 2016.

Y. Wu, V. Tannen, and S. B. Davidson. Priu: A provenance-based approach for incrementally updating regression models. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 447–462, 2020.

C. Yu and W. Yao. Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation*, 46(8):6261–6282, 2017.